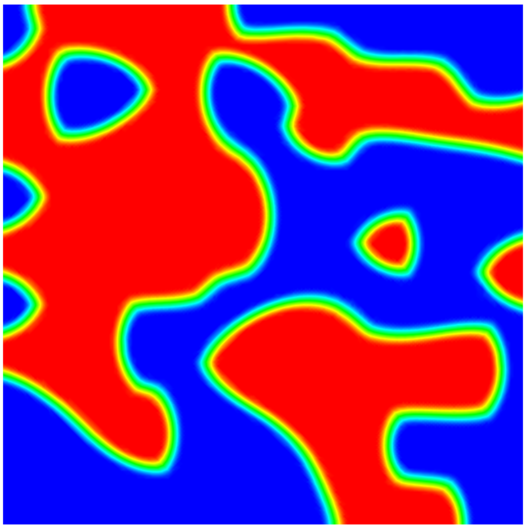
Convex Minimization and Phase Field Models

Inauguraldissertation
zur Erlangung des Grades eines
Doktors der Naturwissenschaften

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin



vorgelegt von
Carsten Gräser

Berlin 2011

Betreuer:	**Prof. Dr. Ralf Kornhuber**	(Freie Universität Berlin)
Erstgutachter:	**Prof. Dr. Ralf Kornhuber**	(Freie Universität Berlin)
Zweitgutachter:	**Prof. Dr. Charles M. Elliott**	(University of Warwick)
Tag der Disputation:	**30. September 2011**	

Contents

1 Introduction

Since the pioneering work of Cahn and Hilliard [29] phase field models have become an increasingly popular tool to model processes involving thin interface layers between almost homogeneous regions. In the original articles they were used to describe separation processes of phases in a physical system [29]. With these first models all phase field models share two characteristics:

- The separation process of phases is driven by a free energy that incorporates a double-well potential with distinct minima for each phase.

- The evolution of the interface between phases is driven by an interfacial energy that penalizes interfaces with high curvature.

This approach has proven to be useful for the simulation of many different processes. While it was originally proposed for the separation of conserved phases it was later also used to model transition processes where phases are transformed into each other [2].

In physical applications the term "phase" can denote different things. In one class of models the phases represent different states of matter [28, 86]. A prototypic example for this class is the Allen–Cahn equation [2]. Another class of models describes separation and solidification processes of binary or multicomponent alloys. For separation processes phases can represent concentrations of different metals in alloys [17]. The prototype for this kind of models is given by the Cahn–Hilliard equation [29, 50]. If solidification processes are modeled the phases typically describe different stable states derived from phase diagrams, while concentrations are described using an additional order parameter [103]. This leads to problems that couple Allen–Cahn and Cahn–Hilliard type equations as considered in [7].

Another important application is the approximation of geometric flows describing the evolution of surfaces (see, e.g., [41]). A well known example is the mean curvature flow, which can be approximated by the zero level set of the solution of an Allen–Cahn equation in the surrounding space [40, 91]. Also, the motion by surface diffusion [31] is the asymptotic limit of a Cahn–Hilliard equation with a degenerate mobility [32].

For all of these applications the selection of the double-well potential plays a crucial role. While smooth potentials have been used successfully for the approximation of geometric flows [13], models describing physical processes often lead to strongly nonlinear or even nonsmooth potentials [17, 103]. Cahn and Hilliard [29] proposed the temperature dependent logarithmic potential. Although this potential is differentiable for positive temperatures it has singular derivatives. For the limiting case of zero temperature the potential degenerates to the obstacle potential, which is no longer differentiable. A common approach is to avoid this situation by replacing the (asymptotically) non-differentiable logarithmic potential by a polynomial one. However, this does not provide a reasonable approximation for small temperatures. Thus, any solution method that is to be applicable for small temperatures must be able to cope with non-differentiable or asymptotically non-differentiable potentials.

While the interfacial energy is often assumed to be isotropic, anisotropic energies are also not uncommon [26, 30, 31, 52, 90]. Non-convex and non-differentiable energies lead to problems where even the unique existence of solutions is problematic. In contrast, convex, differentiable interfacial energies are typically good-natured. Surprisingly many solution methods can only handle isotropic interfacial energies.

The purpose of this thesis is the development of methods for the efficient numerical solution of phase field equations with nonsmooth potentials and anisotropic interfacial energies using finite elements. While this does explicitly include non-differentiable and asymptotically non-differentiable potentials, we will not discuss non-convex and non-differentiable anisotropic interfacial energies. For nonsmooth convex minimization problems obtained by Allen–Cahn type equations efficient adaptive multigrid methods have already been introduced in [62, 72, 76]. The central result of this thesis is the development and analysis of the "Schur Nonsmooth Newton Method" for the solution of nonsmooth nonlinear saddle point problems obtained by finite element discretization of Cahn–Hilliard type equations.

While there are many generic optimization methods for nonlinear saddle point problems, none of them exploits the special structure of the phase field model. We will use this structure to develop methods for the Cahn–Hilliard equation that are comparable to multigrid methods for linear elliptic problems in their efficiency. The general philosophy of the new methods is to use convexity instead of differentiability. By following this idea the methods are inherently robust even for nonsmooth potentials.

The outline of this thesis is as follows. In Chapter 2 we introduce Allen–Cahn and Cahn–Hilliard equations as gradient flows for Ginzburg–Landau energies and give an overview of the present solution theory. Since efficient methods for Allen–Cahn type equations have already been discussed elsewhere [76] we concentrate on the Cahn–Hilliard equation from then on.

Chapter 3 is dedicated to the discussion of finite element discretizations for the Cahn–Hilliard equation. For later use with spatial adaptivity we introduce conforming finite element spaces on nonconforming grids. Afterwards we give a survey of existing finite element discretizations for the Cahn–Hilliard equation. These do in general use uniform grids and are restricted to the isotropic case. We generalize these discretizations to anisotropic equations and time-dependent adaptive grids using Rothe's method. In each time step this leads to a sequence of nonlinear stationary saddle point problems that discretize a continuous saddle point problem. The inherent convex structure allows to show existence and uniqueness of solutions for the discrete as well as for the continuous saddle point problem under reasonable assumptions.

Before we discuss the algebraic solution of these saddle point problems we consider in Chapter 4 the solution of minimization problems for the convex energies associated with the saddle point problems. We deal with these problems here because their solution is a crucial part of the iterative solver that will be developed in Chapter 5 for the saddle point problems. Although efficient multigrid methods for this kind of problem have already been developed for the isotropic case we introduce the new "Truncated Nonsmooth Newton Multigrid" (TNNMG) method. Unlike existing methods the TNNMG method also covers the anisotropic case. Furthermore, we will show that the new approach unifies nonlinear multigrid methods and active set methods.

Chapter 5 is dedicated to the development and analysis of the Schur Nonsmooth Newton method. The corner stone for this solver is an equivalent dual convex minimization problem associated to the nonsmooth nonlinear saddle point problem. The derivative of the energy for this dual problem turns out to be the non-differentiable nonlinear Schur complement of the saddle point problem. For the solution of this dual minimization problem we present the general framework of gradient-related descent methods and extend the known convergence results. After deriving generalized linearizations of the Schur complement we introduce the Schur Nonsmooth Newton method. The convex structure allows to apply the convergence theory for gradient-related descent methods to prove the global convergence of the Schur Nonsmooth Newton method. We will also show that the convergence result is robust with respect to the inexact solution of the linear Newton systems. Since we need to solve a minimization problem for the convex energy of the saddle point problem in each iteration step, we can use the TNNMG method derived in Chapter 4 here. Finally we show that the Schur Nonsmooth Newton method is essentially a globalization of the primal–dual active set method for a subclass of problems where the latter is applicable.

The last ingredient is the construction of adaptive grids in each time step. For this purpose Chapter 6 introduces a hierarchical error estimator for the nonlinear saddle point problems. Following the standard strategy the local contributions of this estimator can be used as local refinement indicators. Besides this, we discuss techniques for the efficient implementation of Rothe's method with time-dependent adaptive grids.

Numerical examples for the introduced methods are presented in Chapter 7. We will especially investigate the mesh independence of the developed Schur Nonsmooth Newton method and the robustness with respect to the temperature.

There is a collection of some useful auxiliary results in Appendix A, a list of symbols in Appendix B, and a list of the key assumptions in Appendix C.

Acknowledgment. Many people supported me during the preparation of this thesis. First of all I would like to thank my supervisor Prof. Dr. Ralf Kornhuber for his constant support, his patience, and his valuable advise. I would also like to thank Prof. Dr. Rolf Krause for getting me into scientific computing, Dr. Heiko Berninger for his clever ideas in functional analysis, Uli Sack for the fruitful collaboration on the implementation, and Dr. Oliver Sander and Dr. Ralf Forster for many interesting discussions and proof-reading. Finally, I am deeply grateful to Juliane and Marjetta for their love and for always reminding me of the beauty of life.

2 Phase Field Models with Nonsmooth Potential

This chapter is a brief introduction to phase field models. In Section 2.1 we present phase field models as gradient flows for Ginzburg–Landau energies. We then give a survey of existence and uniqueness results in Section 2.2.

2.1 Basic Phase Field Models

We start by giving a short motivation of Ginzburg–Landau energies with double-well potentials and derive weak formulations for the evolution equations obtained if gradient flows for these energies are postulated. For a more comprehensive presentation we refer to the monographs of Brokate and Sprekels [25], Eck et al. [48] and the article of Elliott [50].

2.1.1 Anisotropic Ginzburg–Landau Energies

We intend to model phase transition and separation processes in a given spatial domain and a prescribed time interval. To this end let $\Omega \subset \mathbb{R}^d$ with $d = 1, 2, 3$ be a bounded, open, and nonempty set with Lipschitz boundary representing the domain and $[0, T]$ the time interval. The measure of Ω is denoted by $|\Omega|$. Assume that the domain Ω is covered by two phases A and B such that at each time $t \in [0, T]$ in each point $x \in \Omega$ either exactly one phase or a mixture of both phases is present. Depending on the modeled physical process the phases can for example represent different states of matter of a chemical substance, different components of an alloy, or other locally homogeneous states of a system.

In order to describe the presence of pure phases and mixtures we introduce a so-called order parameter u given as function

$$u : \Omega \times [0, T] \to \mathbb{R}.$$

The presence of phases is encoded by

$$u(x,t) \begin{cases} = -1 & \text{if only phase A is present in } (x,t), \\ = 1 & \text{if only phase B is present in } (x,t), \\ \in (-1,1) & \text{if a mixture of A and B is present in } (x,t). \end{cases} \tag{2.1}$$

With this approach $(1 - u)/2$ and $(1 + u)/2$ describe the fractions of phase A and B, respectively. Having this interpretation of u the constraint $u(x, t) \in [-1, 1]$ arises naturally. To simplify notation we will often skip the arguments t or (x, t) if properties or expressions are obviously meant to be written for all t or for all (x, t).

While we will only consider the case of two phases there are also generalizations for $N \geq 2$ phases A_i. In this case one introduces N order parameters given as a function

$$\tilde{u} : \Omega \times [0, T] \to \mathbb{R}^N,$$

where the i-th component $\tilde{u}_i(x, t)$ denotes the fraction of phase A_i in the mixture at (x, t). Then the natural constraints on $\tilde{u}$ are that $\tilde{u}_i(x, t) \in [0, 1]$ and $\sum_{i=1}^N \tilde{u}_i(x, t) = 1$. In the special case of $N = 2$ the representation (2.1) is obtained by setting

$$u = \tilde{u}_2 - \tilde{u}_1,$$

or, for the opposite direction, setting

$$\tilde{u}_1 = \frac{1-u}{2}, \qquad \tilde{u}_2 = \frac{1+u}{2}.$$

Assuming that the phase state at each (x, t) is independent of the state at any other point in space it is natural to postulate a global free energy

$$\psi(u(t)) = \int_\Omega \Psi(u(x, t))dx$$

generated by a local potential Ψ. Since we will introduce evolutions that tend to minimize energies which incorporate ψ it is also possible to incorporate the constraint $u(x, t) \in [-1, 1]$ by using local potentials of the form

$$\Psi : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$$

that take the value ∞ for $u(x, t) \notin [-1, 1]$. In case of transition processes for physical phases or spinodal decomposition of alloys the potential Ψ can be derived and calibrated using phase diagrams that depict which phases or mixtures are stable depending on the temperature θ (see [103]). In general the potential is temperature dependent. Since mixtures tend to be unstable below a critical temperature θ_c the potential Ψ should be some kind of double-well potential that energetically prefers (almost) pure phases in this case. The prototypic example is the logarithmic potential

$$\Psi_\theta(u) = \widehat{\Psi}_\theta(u) - \frac{\theta_c}{2}u^2, \qquad u \in (-1, 1), \tag{2.2}$$

where $\widehat{\Psi}_\theta$ is given by

$$\widehat{\Psi}_\theta(u) = \frac{\theta}{2}\left[(1+u)\ln(1+u) + (1-u)\ln(1-u)\right], \qquad u \in (-1, 1). \tag{2.3}$$

$\widehat{\Psi}_\theta$ and $-\theta_c/2u^2$ are called the convex and concave part of the potential Ψ_θ.

Proposition 2.1. *For $\theta > 0$ the potentials $\Psi_\theta : (-1, 1) \to \mathbb{R}$ given by* (2.2) *and $\widehat{\Psi}_\theta : (-1, 1) \to \mathbb{R}$ given by* (2.3) *have the following properties:*

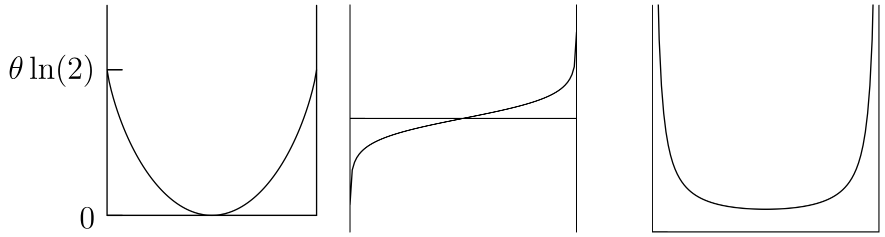


Figure 2.1: Convex part of logarithmic potential: $\widehat{\Psi}_\theta$ (left), $\widehat{\Psi}'_\theta$ (middle), $\widehat{\Psi}''_\theta$ (right).

1. *$\widehat{\Psi}_\theta$ and Ψ_θ are symmetric around $u = 0$.*
2. *$\widehat{\Psi}_\theta$ and Ψ_θ are infinitely differentiable on $(-1, 1)$ with*
$$\widehat{\Psi}'_\theta(u) = \frac{\theta}{2}\ln\left(\frac{1+u}{1-u}\right), \qquad \widehat{\Psi}''_\theta(u) = \frac{\theta}{1-u^2},$$
$$\Psi'_\theta(u) = \widehat{\Psi}'_\theta(u) - \theta_c u, \qquad \Psi''_\theta(u) = \widehat{\Psi}'_\theta(u) - \theta_c.$$
3. *We have $\widehat{\Psi}_\theta(0) = \Psi_\theta(0) = 0$, and for $u \to \pm 1$ we get*
$$\widehat{\Psi}_\theta(u) \to \theta\ln(2), \qquad \widehat{\Psi}'_\theta(u) \to \pm\infty, \qquad \widehat{\Psi}''_\theta(u) \to \infty,$$
$$\Psi_\theta(u) \to \theta\ln(2) - \theta_c/2, \qquad \Psi'_\theta(u) \to \pm\infty, \qquad \Psi''_\theta(u) \to \infty.$$
4. *The logarithmic part $\widehat{\Psi}_\theta(u)$ is strictly convex for all $\theta > 0$.*
5. *For $\theta \geq \theta_c$ the potential Ψ_θ is strictly convex and has a global minimum at $u = 0$.*
6. *For $\theta < \theta_c$ the potential Ψ_θ has a local maximum at $u = 0$ and two global minima in β_θ and $-\beta_\theta$ for some $\beta_\theta \in (0, 1)$. The minima have the property that $\beta_\theta \to 1$ for $\theta \to 0$.*
7. *For $\theta \to 0$ we have $\|\widehat{\Psi}_\theta\|_\infty \to 0$.*

From these properties we instantly get that for all $\theta > 0$ the minimizers of ψ must be contained in $(-1, 1)$ almost everywhere. Hence we will from now on use the natural extensions of Ψ_θ and $\widehat{\Psi}_\theta$ to $\mathbb{R}$ given by $\Psi_\theta(u) = \widehat{\Psi}_\theta(u) - \frac{\theta_c}{2}u^2$ and

$$\widehat{\Psi}_\theta(u) = \begin{cases} \frac{\theta}{2}\left[(1+u)\ln(1+u) + (1-u)\ln(1-u)\right] & \text{if } u \in (-1, 1), \\ \theta\ln(2) & \text{if } |u| = 1, \\ \infty & \text{else,} \end{cases}$$

respectively. Since the definition remains the same in $(-1, 1)$ we use the same symbols.

In view of the limiting properties in Proposition 2.1 a formal limit $\Psi_0(u) = \widehat{\Psi}_0(u) - \frac{\theta_c}{2}u^2$ of Ψ_θ for $\theta \to 0$ should have minima in -1 and 1 and $\widehat{\Psi}_0(u)$ should be zero for

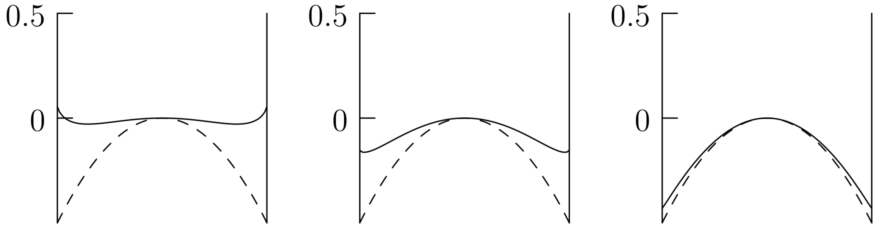


Figure 2.2: Logarithmic potential Ψ_θ (solid) in comparison with Ψ_0 (dashed) for $\theta_c = 1$ and $\theta = 0.8$ (left), $\theta = 0.5$ (middle), $\theta = 0.1$ (right).

$u \in (-1, 1)$. Furthermore, it should also ensure that its minima are contained in the domain $[-1, 1]$ even if another smooth functional is added. Under these conditions the only possible extension is the so-called obstacle potential $\Psi_0 : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ given by

$$\Psi_0(u) = \widehat{\Psi}_0(u) - \frac{\theta_c}{2}u^2, \qquad \widehat{\Psi}_0(u) = \chi_{[-1,1]}(u).$$

Here we have used the indicator functional $\chi_K : V \to \mathbb{R} \cup \{\infty\}$ for a subset $K \subset V$ of a vector space V defined by

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

Although the logarithmic potential Ψ_θ is smooth on its domain $(-1, 1)$ for $\theta > 0$ it does rapidly degenerate to the nonsmooth obstacle potential for $\theta \to 0$. Due to this fact the nonlinearity can be regarded as de facto nonsmooth even for values $\theta \in (0, \theta_c)$ that are close to θ_c (see Figure 2.2). Furthermore, the unbounded derivatives near -1 and $+1$ are numerically challenging even for values of θ very close to θ_c. Thus many authors replace the logarithmic potential by a quartic approximation that is only valid for $\theta \approx \theta_c$. Since we do not want to impose such a restriction we will always consider the global logarithmic potential

$$\psi_\theta(u) = \int_\Omega \Psi_\theta(u(x))dx = \widehat{\psi}_\theta(u) - \frac{\theta_c}{2}\int_\Omega u(x)^2 dx, \qquad \widehat{\psi}_\theta(u) = \int_\Omega \widehat{\Psi}_\theta(u(x))dx$$

for all $\theta \geq 0$ including the special limiting case of the obstacle potential ($\theta = 0$).

In more general situations the assumption of no interaction between neighboring points does not hold. For example in case of the spinodal decomposition of alloys larger areas consisting of only one phase are observed and these areas tend to minimize the curvature of their boundary (see Figure 2.3. In view of this observation it is reasonable to introduce a nonlocal surface term that penalizes high curvatures. The simplest form of such a term is $\overline{\gamma}\|\nabla u\|^2$ for a fixed constant $\overline{\gamma} > 0$. However, this term does only model isotropic behavior.

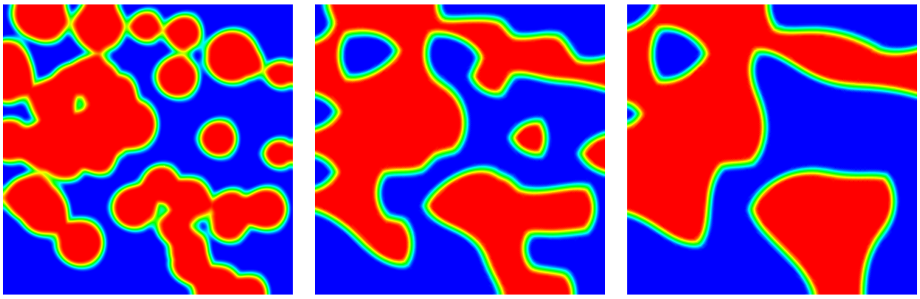


Figure 2.3: Minimization of interface curvature during anisotropic coarsening.

In order to also capture anisotropic effects where certain spatial directions are preferred we will use the term $\gamma(\nabla u)^2$ for a function

$$\gamma : \mathbb{R}^d \to \mathbb{R}.$$

The squared function γ^2 should, however, behave like a quadratic functional in the following sense.

(A1) The function $\gamma : \mathbb{R}^d \to \mathbb{R}$ is positive 1-homogeneous, i.e.,

$$\gamma(x) \geq 0, \qquad \gamma(\lambda x) = \lambda\gamma(x), \qquad \forall x \in \mathbb{R}^d, \lambda > 0,$$

definite, i.e.,

$$\gamma(x) = 0 \qquad \Rightarrow \qquad x = 0,$$

and twice continuously differentiable on $\mathbb{R}^d \setminus \{0\}$.

Functions satisfying (A1) naturally induce similar scaling properties for their derivatives. From now on $\nabla^2 f$ will always denote the Hessian matrix of a functional $f : \mathbb{R}^d \supset M \to \mathbb{R}$.

Lemma 2.1. *Let $\gamma : \mathbb{R}^d \to \mathbb{R}$ satisfy (A1). Then for all $x \in \mathbb{R}^d \setminus \{0\}$ and $\lambda > 0$ we have*

$$\begin{aligned} (\nabla\gamma)(\lambda x) &= \nabla\gamma(x), & \lambda\nabla^2\gamma(\lambda x) &= \nabla^2\gamma(x), \\ \langle\nabla\gamma(x), x\rangle &= \gamma(x) > 0, & \left\langle\nabla^2\gamma(x)x, x\right\rangle &= 0. \end{aligned}$$

Proof. Let $x \in \mathbb{R}^d \setminus \{0\}$. For $\lambda > 0$ define $g_\lambda(x) := \gamma(\lambda x)$ and $h_\lambda(x) := (\nabla\gamma)(\lambda x)$. Then we have $g_\lambda(x) = \lambda\gamma(x)$ and thus

$$\lambda(\nabla\gamma(x)) = \nabla g_\lambda(x) = (\nabla\gamma)(\lambda x)\lambda.$$

Using this we get $h_\lambda(x) = \nabla\gamma(x)$ and

$$\nabla^2\gamma(x) = Dh_\gamma(x) = \nabla^2\gamma(\lambda x)\lambda.$$

Furthermore,

$$\langle \nabla\gamma(x), x\rangle = \lim_{h\to 0} \frac{\gamma(x+hx)-\gamma(x)}{h} = \lim_{h\to 0} \frac{(1+h)\gamma(x)-\gamma(x)}{h} = \gamma(x)$$

and for $f : \mathbb{R} \to \mathbb{R}, f(t) = \gamma(tx)$ the chain rule and the scaling property of $\nabla\gamma$ imply

$$f'(t) = \langle (\nabla\gamma)(tx), x\rangle = \langle \nabla\gamma(x), x\rangle = \gamma(x), \qquad f''(t) = \left\langle \nabla^2\gamma(tx)x, x\right\rangle = 0.$$

□

The scaling property of γ also directly implies certain properties of γ^2:

Lemma 2.2. *Assume that $\gamma : \mathbb{R}^d \to \mathbb{R}$ satisfies (A1). Then we have:*

1. *γ^2 is continuously differentiable on $\mathbb{R}^d$ with*

$$\nabla(\gamma^2)(x) = 2\gamma(x)(\nabla\gamma)(x) \qquad \forall x \in \mathbb{R}^d \setminus \{0\}, \qquad \nabla(\gamma^2)(0) = 0.$$

2. *γ^2 is twice continuously differentiable on $\mathbb{R}^d \setminus \{0\}$ with*

$$\nabla^2(\gamma^2)(x) = 2\left(\nabla\gamma(x)^T\nabla\gamma(x) + \gamma(x)\nabla^2\gamma(x)\right) \qquad \forall x \in \mathbb{R} \setminus \{0\}.$$

3. *γ^2, $\nabla(\gamma^2)$, and $\nabla^2(\gamma^2)$ have the scaling properties*

$$\begin{aligned} \gamma^2(\lambda x) &= \lambda^2\gamma^2(x) && \forall x \in \mathbb{R}^d, \lambda > 0,\\ \nabla(\gamma^2)(\lambda x) &= \lambda\nabla(\gamma^2)(x) && \forall x \in \mathbb{R}^d, \lambda > 0,\\ \nabla^2(\gamma^2)(\lambda x) &= \nabla^2(\gamma^2)(x) && \forall x \in \mathbb{R}^d \setminus \{0\}, \lambda > 0. \end{aligned}$$

4. *There is a constant $\overline{H}_{\gamma^2} > 0$ such that*

$$\left\langle \nabla^2(\gamma^2)(x)y, y\right\rangle \le \left\langle \overline{H}_{\gamma^2}y, y\right\rangle, \qquad \forall x \in \mathbb{R}^d \setminus \{0\}, y \in \mathbb{R}^d.$$

5. *$\nabla(\gamma^2)$ is Lipschitz continuous with Lipschitz constant $\overline{H}_{\gamma^2}$.*

6. *γ^2 is coercive, i.e., $\gamma(x)^2 \ge \gamma^2_{\min}\|x\|^2$ with $\gamma^2_{\min} > 0$.*

7. *For $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^d \setminus \{0\}$ we have*

$$\left\langle \nabla(\gamma^2)(x), x\right\rangle = 2\gamma^2(x), \qquad \left\langle \nabla^2(\gamma^2)(y)y, y\right\rangle = 2\gamma^2(y) > 0.$$

Proof. Existence, continuity, and the representations of $\nabla(\gamma^2)$ and $\nabla^2(\gamma^2)$ on $\mathbb{R}^d \setminus \{0\}$ are provided by the chain rule, while $\nabla(\gamma^2)(0) = 0$ follows from the scaling property and boundedness of γ on $S^{d-1} := \{z \in \mathbb{R}^d : \|z\| = 1\}$. By the boundedness of γ and $\nabla\gamma$ on S^{d-1} and their scaling properties we have for $x \in \mathbb{R}^d \setminus \{0\}$

$$\|\nabla(\gamma^2)(x)\| = \|2\gamma(x)(\nabla\gamma)(x)\| \le C\|x\|,$$

which implies continuity of $\nabla(\gamma^2)$ in 0 and thus on the whole set $\mathbb{R}^d$.

The scaling property of γ^2 follows directly from the corresponding property of γ, and those of $\nabla(\gamma^2)$ and $\nabla^2(\gamma^2)$ from Lemma 2.1.

By continuity the mapping $(x, y) \mapsto \langle\nabla^2(\gamma^2)(x)y, y\rangle$ is bounded on $S^{d-1} \times S^{d-1}$. Hence there is a constant $\overline{H}_{\gamma^2} \ge 0$ with

$$\langle\nabla^2(\gamma^2)(x)y, y\rangle \le \overline{H}_{\gamma^2} \qquad \forall x, y \in S^{d-1}.$$

Now boundedness of $\nabla^2(\gamma^2)$ follows from its scaling property.

To show Lipschitz continuity of $\nabla(\gamma^2)$ let $x, y \in \mathbb{R}^n$. First we consider $0 \notin \operatorname{co}\{x, y\}$, where $\operatorname{co} M$ denotes the convex hull of M. Then there is an open convex set U with $x, y \in U$ and $0 \notin U$. Hence $\nabla(\gamma^2)$ is continuously differentiable on U and thus the mean value theorem and the scaling property, symmetry, and the bound of $\nabla^2(\gamma^2)$ provide

$$\|\nabla(\gamma^2)(y) - \nabla\gamma^2(x)\| \le \int_0^1 \|\nabla^2(\gamma^2)(x + t(y - x))\| \, \|y - x\| dt \le \overline{H}_{\gamma^2}\|y - x\|.$$

Now let $0 \in \operatorname{co}\{x, y\}$. Then for $\epsilon > 0$ the above provides

$$\|\nabla(\gamma^2)(y) - \nabla(\gamma^2)(\epsilon y) + \nabla(\gamma^2)(\epsilon x) - \nabla(\gamma^2)(x)\| \le \overline{H}_{\gamma^2}\|y - \epsilon y\| + \overline{H}_{\gamma^2}\|x - \epsilon x\|.$$

Taking the limit $\epsilon \to 0$ and using the continuity of $\nabla(\gamma^2)$ we get

$$\|\nabla(\gamma^2)(y) - \nabla(\gamma^2)(x)\| \le \overline{H}_{\gamma^2}\|x\| + \overline{H}_{\gamma^2}\|y\| = \overline{H}_{\gamma^2}\|x - y\|.$$

Coercivity follows from continuity since $\gamma_{\min}^2 := \min_{z \in S^{d-1}} \gamma(z)^2 > 0$ and thus

$$\gamma(x)^2 = \|x\|^2\gamma\left(\frac{x}{\|x\|}\right)^2 \ge \gamma_{\min}^2\|x\|^2.$$

Finally the representations of $\langle\nabla(\gamma^2)(x), x\rangle$ and $\langle\nabla^2(\gamma^2)(y)y, y\rangle$ follow from Lemma 2.1. □

Nonconvex anisotropies allowed by Assumption (A1) may still lead to ill-posed equations that require regularization techniques (see, e.g., [105, 109]). In order to avoid these problems we introduce a stronger assumption to ensure the strong convexity of γ^2.

(A2) The function $\gamma : \mathbb{R}^d \to \mathbb{R}$ satisfies (A1) and

$$\nabla^2(\gamma^2)(x) = 2\left(\nabla\gamma(x)^T\nabla\gamma(x) + \gamma(x)\nabla^2\gamma(x)\right)$$

is symmetric positive definite for all $x \neq 0$.

Sometimes it might be easier to check the following stronger assumption for γ.

(A2') The function $\gamma : \mathbb{R}^d \to \mathbb{R}$ is convex and satisfies (A1) and $\ker \nabla^2\gamma(x) = \operatorname{span}\{x\}$ for all $x \neq 0$.

Note that (A1) already implies that γ is linear along rays originating in 0 and hence $\operatorname{span}\{x\} \subset \ker \nabla^2\gamma(x)$. Thus assumption (A2) essentially adds the requirement that γ is strictly convex in directions tangential to the unit sphere.

Lemma 2.3. *(A2') implies (A2) and for convex γ (A2) implies (A2').*

Proof. Let γ satisfy (A2') and $x, y \in \mathbb{R}^d \setminus \{0\}$. If $y \notin \ker \nabla^2\gamma(x)$ we have

$$\left\langle \nabla^2(\gamma^2)(x)y, y\right\rangle \geq 2\left\langle \nabla^2\gamma(x)y, y\right\rangle > 0,$$

whereas for $y \in \ker \nabla^2\gamma(x)$ Lemma 2.2 implies

$$\left\langle \nabla^2(\gamma^2)(x)y, y\right\rangle = \frac{\|y\|^2}{\|x\|^2}\left\langle \nabla^2(\gamma^2)(x)x, x\right\rangle = \frac{\|y\|^2}{\|x\|^2}2\gamma^2(x) > 0.$$

Now let γ be convex, satisfying (A2), and $x \in \mathbb{R}^d \setminus \{0\}$. Then

$$d = \operatorname{rank} \nabla^2(\gamma^2)(x) \leq \operatorname{rank}\left(\nabla\gamma(x)^T\nabla\gamma(x)\right) + \operatorname{rank}\nabla^2\gamma(x) = 1 + \operatorname{rank}\nabla^2\gamma(x)$$

and hence $\operatorname{rank}\nabla^2\gamma(x) \geq 1 - d$. Together with $\operatorname{span}\{x\} \subset \ker \nabla^2\gamma(x)$ this gives (A2'). □

Lemma 2.4. *Assume that $\gamma : \mathbb{R}^d \to \mathbb{R}$ satisfies (A2). Then $\gamma^2 : \mathbb{R}^d \to \mathbb{R}$ satisfies:*

1. *The Hessian of γ^2 is uniformly bounded from below, i.e., there is $\underline{H}_{\gamma^2} > 0$ such that*

$$\underline{H}_{\gamma^2}\langle y, y\rangle \leq \left\langle \nabla^2(\gamma^2)(x)y, y\right\rangle \qquad \forall x \in \mathbb{R}^d \setminus \{0\}, y \in \mathbb{R}^d.$$

2. *$\nabla(\gamma^2)$ is strongly monotone, i.e.,*

$$\left\langle \nabla(\gamma^2)(x) - \nabla(\gamma^2)(y), x - y\right\rangle \geq \underline{H}_{\gamma^2}\|x - y\|^2 \qquad \forall x, y \in \mathbb{R}^m. \tag{2.4}$$

3. *γ^2 is strongly convex, i.e.,*

$$\gamma^2(\lambda x + (1-\lambda)y) \leq \lambda\gamma^2(x) + (1-\lambda)\gamma^2(y) - \lambda(1-\lambda)\frac{\underline{H}_{\gamma^2}}{2}\|x - y\|^2 \qquad \forall \lambda \in [0, 1].$$

4. γ^2 is coercive with $\gamma(x)^2 \geq \frac{\underline{H}_{\gamma^2}}{2}\|x\|^2$, i.e., $\gamma_{\min}^2 \geq \frac{\underline{H}_{\gamma^2}}{2} > 0$.

Proof. Due to the scaling property of $\nabla^2(\gamma^2)$ (see Lemma 2.2) we only have to show boundedness for $x \in S^{d-1}$. Since the continuous mapping $(x, y) \mapsto \langle \nabla^2(\gamma^2)(x)y, y\rangle$ does only take positive values on $S^{d-1} \times S^{d-1}$ the infimum $\underline{H}_{\gamma^2}$ of those values on the compact set $S^{d-1} \times S^{d-1}$ is also positive.

Since $\nabla^2(\gamma^2)$ is symmetric positive semidefinite and bounded from below on $\mathbb{R}^d \setminus \{0\}$ we can use Lemma A.3 in the appendix to obtain strong monotonicity of $\nabla(\gamma^2)$ and hence by Lemma A.1 in the appendix strong convexity of γ^2.

Using Lemma 2.2 and the strong monotonicity we get

$$2\gamma^2(x) = \langle \nabla(\gamma^2)(x), x\rangle \geq \underline{H}_{\gamma^2}\|x\|^2.$$

□

Obviously any norm that is smooth enough satisfies (A1). However, in contrast to norms, functions satisfying (A1) need not be symmetric with respect to the origin. This is for example the case for the following anisotropy function introduced by Kobayashi [68].

Example 2.1. *Let $k \in \mathbb{N}$ and $\bar{a} > 0$. For $\xi \in \mathbb{R}^2$ let $\beta(\xi) \in [0, 2\pi]$ denote the angle between the positive x-axis and ξ. Then the anisotropy function*

$$\gamma(x) = \left[1 + \bar{a}\cos(k\beta(x))\right]\|x\|$$

is positive 1-homogeneous, definite and twice continuously differentiable for $x \neq 0$.

If additionally $\bar{a} < 1/(k^2 - 1)$ holds true $\nabla^2(\gamma^2)$ is positive definite (see [26]) and γ satisfies (A2). Note that the integer k denotes the number of directions where the "unit sphere" $\{x \in \mathbb{R}^2 : \gamma(x) = 1\}$ of γ is deformed compared to S^{d-1}. Furthermore, γ is symmetric with respect to rotations by $2\pi/k$.

If the anisotropy is a non-differentiable norm it may still be possible to approximate it by a smooth function. Such an approximation for the 1-norm is for example given by γ_ϵ in the following.

Example 2.2. *For $\epsilon > 0$ the functional $\gamma_\epsilon : \mathbb{R}^d \to \mathbb{R}$ given by*

$$\gamma_\epsilon(x) = \sum_{i=1}^{d} (x_i^2 + \epsilon\|x\|^2)^{\frac{1}{2}}$$

satisfies (A2). For $\epsilon \to 0$ we have $\gamma_\epsilon(x) \to \sum_{i=1}^{d} |x_i|$.

For a function γ satisfying (A2) we now define the anisotropic Ginzburg–Landau free energy with logarithmic potential

$$E(u) := \psi_\theta(u) + \frac{1}{2}\int_\Omega \gamma(\nabla u(x))^2 dx. \tag{2.5}$$

Note that for $\gamma(\cdot)^2 = \overline{\gamma}\|\cdot\|^2$ this is the standard isotropic Ginzburg–Landau free energy

$$E(u) = \psi_\theta(u) + \frac{1}{2}\int_\Omega \overline{\gamma}\|\nabla u(x)\|^2 dx.$$

2.1.2 Gradient Flows for Ginzburg–Landau Energies

Different phase field models with or without mass conservation can be derived from the Ginzburg–Landau energy by postulating gradient flows with respect to suitable norms (see, e.g., [30]). In order to define these gradients flows we first introduce the appropriate function spaces.

Definition 2.1. *We introduce the following notation for function spaces, norms, and products:*

1. *For vectors in $x, y \in \mathbb{R}^n$ the Euclidean inner product and norm are denoted by $\langle x, y\rangle$ and $\|x\|$, respectively. For a matrix M the induced bilinear form is denoted by $\langle x, y\rangle_M = \langle Mx, y\rangle$. If M is symmetric positive semidefinite the induced semi-norm is $\|x\|_M = \sqrt{\langle x, y\rangle_M}$.*

2. *$C^k(\overline{\Omega})$ is the space of all k-times continuously differentiable functions such that all partial derivatives up to order k are bounded. Its norm is denoted by $\|\cdot\|_{\infty,k}$.*

3. *$C_0^k(\overline{\Omega})$ is the subspace of all functions $u \in C^k(\overline{\Omega})$ such that $u|_{\partial\Omega} = 0$.*

4. *For $p > 0$ the Lebesgue space $L^p(\Omega)$ is the space of all measurable functions $v : \Omega \to \mathbb{R}$ such that $|v(\cdot)|^p$ is integrable. Its norm is given by*

$$\|v\|_{L^p(\Omega)} = \left(\int_\Omega |v(x)|^p \, dx\right)^{\frac{1}{p}}.$$

The norm of $L^2(\Omega)$ is denoted by $\|\cdot\|_0$ and induced by the inner product

$$(v, w) = \int_\Omega v(x)w(x)\, dx.$$

5. *The Sobolev space $H^k(\Omega) \subset L^2(\Omega)$, $k \geq 0$ is the subspace of all functions having weak partial derivatives up to order k in $L^2(\Omega)$. Its norm is denoted by $\|\cdot\|_k$. Furthermore, the $H^1(\Omega)$-semi-norm $\sqrt{(\nabla\cdot, \nabla\cdot)}$ is denoted by $|\cdot|_1$.*

6. *The Sobolev space $H_0^k(\Omega) \subset H^k(\Omega)$ is the closure of $C_0^k(\overline{\Omega})$ in $H^k(\Omega)$ with respect to the norm $\|\cdot\|_k$.*

7. *For a normed space V the natural norm is denoted by $\|\cdot\|_V$.*

8. *For a pre-Hilbert space H the natural inner product is denoted by $(\cdot, \cdot)_H$.*

9. *The dual space of a normed space V is denoted by V'. The dual paring of $x \in V$ and $y \in V'$ is denoted by $y(x) = \langle y, x\rangle$.*

10. *For a Hilbert space H, $T > 0$, and $p > 0$ the Bochner space $L^p(0,T;H)$ is the space of all weakly measurable functions $v : (0,T) \to H$ such that $\int_0^T \|v(t)\|^p dt$ is finite. Its norm is given by*
$$\|v\|_{L^p(0,T;H)} = \left(\int_0^T \|v(t)\|_H^p \, dt \right)^{\frac{1}{p}}.$$
For $p = 2$ this is induced by the inner product
$$(v,w)_{L^2(0,T;H)} = \int_0^T (v(t), w(t))_H \, dt.$$
For a precise definition of weakly measurable and Bochner integrable functions we refer to Wloka [110].

11. *For a Hilbert space H and $T > 0$ the Bochner Sobolev space $W(0,T;H)$ is the subspace of all $v \in L^2(0,T;H)$ with weak time derivative $\frac{dv}{dt}$ in $L^2(0,T;H')$. Its norm is induced by the scalar product*
$$(v,w)_{W(0,T;H)} = (v,w)_{L^2(0,T;H)} + \left(\frac{dv}{dt}, \frac{dw}{dt} \right)_{L^2(0,T;H')}.$$
For a precise definition of the weak time derivative we refer to Wloka [110] and Dautray and Lions [38].

Before postulating gradient flows we compute the gradient of the anisotropic Ginzburg–Landau free energy E at some u. To this end define the nonlinear operator $\mathcal{F}_\gamma : H^1(\Omega) \to (H^1(\Omega))'$ as
$$\langle \mathcal{F}_\gamma(w), v \rangle := \int_\Omega \gamma(\nabla w(x)) \, \langle (\nabla\gamma)(\nabla w(x)), \nabla v(x) \rangle \, dx.$$
This operators turns out to be the derivative of the smooth part of the Ginzburg–Landau energy E. For the isotropic functional $\gamma(x)^2 = \overline{\gamma}\|x\|^2$ with a fixed constant $\overline{\gamma} > 0$ the operator $\mathcal{F}_\gamma$ becomes linear and is given by the bilinear form
$$\langle \mathcal{F}_\gamma w, v \rangle = \overline{\gamma} \, (\nabla w, \nabla v) \, .$$

Lemma 2.5. *Assume that $\gamma : \mathbb{R}^d \to \mathbb{R}$ satisfies (A2). Then the functional $J_\gamma : H^1(\Omega) \to \mathbb{R}$, defined by*
$$J_\gamma(v) := \int_\Omega \frac{1}{2} \gamma(\nabla v(x))^2 dx,$$
is convex, Fréchet differentiable with $\nabla J_\gamma = \mathcal{F}_\gamma$, and continuous on $H^1(\Omega)$.

Proof. By Lemma 2.2 and Lemma 2.4 the functional $f : \mathbb{R}^d \to \mathbb{R}$ with $f(x) = \frac{1}{2}\gamma^2(x)$ is strongly convex and differentiable with the Lipschitz continuous derivative $\nabla f(x) = \gamma(x)\nabla\gamma(x)$. Hence

$$\begin{aligned}\frac{1}{h}\Big(f(x+hy) - f(x)\Big) - \langle \nabla f(x), y\rangle &= \frac{1}{h}\int_0^h \langle \nabla f(x+ty) - \nabla f(x), y\rangle \, dt \\ &\leq \frac{L}{h}\int_0^h t\|y\|^2 dt = \frac{L}{2}h\|y\|^2.\end{aligned}$$

Now let $w, v \in H^1(\Omega)$. Then

$$\left| \frac{1}{h}\Big(J_\gamma(w+hv) - J_\gamma(w)\Big) - \langle \mathcal{F}(w), v\rangle \right| \leq \frac{L}{2}h\,(\nabla v, \nabla v)\,.$$

Since this goes to zero uniformly with respect to $\|v\|_1$ we have shown Fréchet differentiability. Convexity and continuity of J_γ follow directly from convexity of γ^2 and differentiability, respectively. □

For the introduction of gradient flows we first consider the case $\theta > 0$. Then E is differentiable and its gradient at some $u \in H^1(\Omega)$ with $-1 < u < 1$ a.e. (almost everywhere) is given by the functional

$$\nabla E(u) = \int_\Omega \Psi'_\theta(u(x))(\cdot)\, dx + \langle \mathcal{F}_\gamma(w), \cdot\rangle\,,$$

or, equivalently,

$$\langle \nabla E(u), v\rangle = \int_\Omega \Psi'_\theta(u(x))v(x)\, dx + \langle \mathcal{F}_\gamma(w), v\rangle \qquad \forall v \in H^1(\Omega).$$

In the case $\theta = 0$, where Ψ_θ is the obstacle potential, E is no longer differentiable in the classical sense. To overcome the lack of smoothness we recall the notion of the subdifferential (see, e.g., [49]).

Definition 2.2. *Let V be a normed space and $\phi : V \to \mathbb{R} \cup \{\infty\}$ be convex. Then for $x \in V$ the subdifferential $\partial\phi(x) \subset V'$ is the set of all $v \in V'$ such that*

$$\phi(x) + \langle v, y - x\rangle \leq \phi(y) \qquad \forall y \in V.$$

From the definition it is obvious that x minimizes ϕ if and only if $0 \in \partial\phi(x)$. Using this definition we can compute the subdifferential of E at $\theta = 0$ (see [49, Chapter I, Proposition 5.6])

$$\partial E(u) = \partial\psi_\theta(u)(\cdot) + \langle \mathcal{F}_\gamma(w), \cdot\rangle\,.$$

Different phase field models are now obtained as gradient flows for this functional. First consider the L^2 gradient flow

$$\frac{du}{dt} \in -\partial E(u).$$

If u and E are sufficiently smooth this can be interpreted as $\frac{du}{dt}$ being the L^2 projection of $-\nabla E(u)$ to $L^2(\Omega)$, i.e.,

$$\frac{du}{dt} = \operatorname*{arg\,min}_{v \in L^2(\Omega)} \left(\frac{1}{2} \|v\|^2_{L^2(\Omega)} - \langle -\nabla E(u), v \rangle \right),$$

and is hence called an L^2 gradient flow for E. Complemented by the natural boundary conditions for the anisotropic differential operator (using the outer unit normal $n(x)$ of Ω at $x \in \partial\Omega$)

$$\gamma(\nabla u(x)) \big\langle n(x), (\nabla\gamma)(\nabla u(x)) \big\rangle = 0 \qquad \forall x \in \partial\Omega, \quad t > 0,$$

this leads to the anisotropic Allen–Cahn equation. For given initial data $u_0 \in H^1(\Omega)$ the weak formulation of this partial differential equation is given by the following variational inequality problem:

Problem 2.1. *Find $u \in W(0,T,H^1(\Omega))$ such that $u(0) = u_0$ a.e. in Ω and*

$$\left\langle \frac{du}{dt} + \mathcal{F}_\gamma(u), v - u \right\rangle + \psi_\theta(v) - \psi_\theta(u) \geq 0 \qquad \forall v \in H^1(\Omega), \qquad \text{a.e. in } (0,T).$$

Now we want to introduce a conservative flow with the mass conservation property

$$\int_\Omega u(x,t)\,dx = \int_\Omega u(x,0)\,dx \qquad \forall t > 0. \tag{2.6}$$

In contrast to the Allen–Cahn equation we use a H^{-1}-like gradient flow. To this end we introduce the space

$$H_0 = \left\{ v \in H^1(\Omega) : \int_\Omega v(x)\,dx = 0 \right\} \tag{2.7}$$

with the natural inner product $(\nabla\cdot, \nabla\cdot)$. Then the Riesz isomorphism $R_{H_0} : H_0 \to (H_0)'$ is given by

$$(\nabla w, \nabla v) = \langle R_{H_0} w, v \rangle \qquad \forall v, w \in H_0.$$

For smooth u and E we can now postulate the $(H_0)'$ gradient flow

$$\frac{du}{dt} = \operatorname*{arg\,min}_{v \in (H_0)'} \left(\frac{1}{2} \|v\|^2_{(H_0)'} - \langle -\nabla E(u), v \rangle + \lambda(1, v) \right) \tag{2.8}$$

subject to the mass conservation constraint (2.6). Here $\lambda \in \mathbb{R}$ denotes the Lagrangian multiplier for the constraint. Using the Riesz isomorphism we can rewrite this also for nonsmooth E as

$$R^{-1}_{H_0} \frac{du}{dt} \in -\partial E(u) - \lambda, \qquad \int_\Omega \frac{du}{dt}(x)\,dx = 0.$$

Defining the so-called chemical potential $w = R^{-1}_{H_0} \frac{du}{dt} + \lambda$ we get the system

$$w \in \partial E(u), \qquad R_{H_0}(w + \lambda) = -\frac{du}{dt}.$$

A weak formulation for this inclusion is given by the variational inequality problem:

Problem 2.2. *Find* $u \in W(0,T;H^1(\Omega))$ *and* $w \in L^2(0,T;H^1(\Omega))$ *such that* $u(0) = u_0$ *a.e. in* Ω *and*

$$\begin{aligned} \langle \mathcal{F}_\gamma(u), v-u\rangle - (w, v-u) + \psi_\theta(v) - \psi_\theta(u) &\geq 0 \qquad \forall v \in H^1(\Omega), \qquad \textit{a.e. in } (0,T), \\ \left\langle \frac{du}{dt}, v\right\rangle + (\nabla w, \nabla v) &= 0 \qquad \forall v \in H^1(\Omega), \qquad \textit{a.e. in } (0,T). \end{aligned}$$

This problem represents an anisotropic version of the Cahn–Hilliard equation. For isotropic γ it reduces to the well-known Cahn–Hilliard equations with logarithmic potential for $\theta > 0$ and obstacle potential for $\theta = 0$. Note that mass conservation is incorporated by testing the second equation with all $v \in H^1(\Omega)$ instead of $v \in H_0$. This equation can be regarded as the sum of (2.6) tested with all constants and (2.8).

Obviously solutions u of Problem 2.1 and Problem 2.2 cannot exceed the interval $[-1,1]$ on a set with positive measure, i.e., they must stay in the set

$$\mathcal{K} := \{v \in H^1(\Omega) : |v| \leq 1 \text{ a.e. in } \Omega\}, \tag{2.9}$$

which is just the domain

$$\operatorname{dom} \widehat{\psi}_\theta := \{v \in H^1(\Omega) : \widehat{\psi}_\theta(v) < \infty\}$$

of $\widehat{\psi}_\theta$.

2.2 Solution Theory

In this section we consider the existence and uniqueness of solutions to the variational problems introduced in the previous section. While we will not develop a solution theory for the general case of Problem 2.1 and Problem 2.2 including anisotropy and the logarithmic or obstacle potential, we give a short overview of the known results for certain special cases. The results will be adapted to the presented framework by changing the notation and rescaling parameters where necessary.

2.2.1 Allen–Cahn Equation

First we consider the Allen–Cahn type equations arising from the L^2 gradient flow for E. In Chen and Elliott [33] the obstacle potential, i.e., $\theta = 0$, with an isotropic $\gamma : \mathbb{R}^d \to \mathbb{R}$ of the form

$$\gamma(x)^2 = \overline{\gamma}\|x\|^2$$

with some constant $\overline{\gamma} > 0$ was discussed. For this case Problem 2.1 can be written as the following parabolic variational inequality.

Problem 2.3. *Find* $u \in W(0,T;H^1(\Omega))$ *such that* $u(0) = u_0$ *a.e. in* Ω*,* $u(t) \in \mathcal{K}$ *and*

$$\left\langle \frac{du}{dt}, v-u\right\rangle + \overline{\gamma}\,(\nabla u, \nabla(v-u)) \geq \theta_c\,(u, v-u) \qquad \forall v \in \mathcal{K}, \qquad \textit{a.e. in } (0,T).$$

For this problem the authors prove the following existence and stability result:

Theorem 2.1. *Let $u_0 \in L^\infty(\Omega)$ and $\|u_0\|_\infty \leq 1$. Then there is a unique solution u of Problem 2.3 that satisfies $u(t) \in \mathcal{K}$ a.e. in $(0,T)$. Furthermore, u satisfies $u \in C(0,T;L^2(\Omega))$ and*

$$E(u(t)) + \int_{t'}^{t} \left\| \frac{du}{dt}(s) \right\| ds = E(u(t'))$$

for all $t, t' \in [0,T]$ with $t' < t$. Hence E is a Lyapunov functional for Problem 2.3.

The anisotropic case was considered by Burman and Rappaz [26]. There a coupled system where the concentration and the phase field variable do not coincide is studied.

Theorem 2.2. *Consider $\gamma : \mathbb{R}^2 \to \mathbb{R}$ as in Example 2.1 with $\bar{a} < 1/(k^2-1)$ and Lipschitz continuous functions $D_1 : \mathbb{R} \to \mathbb{R}$ and $D_2, S : \mathbb{R}^2 \to \mathbb{R}$ with $0 < D_s < D_1(r) \leq D_l$ $\forall r \in \mathbb{R}$. Let $u_0 \in L^2(\Omega)$ and $c_0 \in L^2(\Omega)$. Then there are $u, c \in W(0,T;H^1(\Omega))$ that satisfy*

$$\left\langle \frac{du}{dt}, v \right\rangle + \langle \mathcal{F}_\gamma(u), v\rangle - (S(c,u), v) = 0 \quad \forall v \in H^1(\Omega), \quad a.e.\ in\ (0,T),$$

$$\left\langle \frac{dc}{dt}, v \right\rangle + \int_\Omega \langle D_1(u)\nabla c + D_2(c,u)\nabla u, \nabla v\rangle \, dx = 0 \quad \forall v \in H^1(\Omega), \quad a.e.\ in\ (0,T).$$

Furthermore, if $u_0 \in H^1(\Omega)$ the solution satisfies

$$u \in L^\infty(0,T;H^1(\Omega)) \cap H^1(\Omega \times (0,T)).$$

If the parameter functions are chosen as $D_1(u) = 1$, $D_2(c,u) = 0$, and $S(c,u) = -\Psi'(u)$ the equations in Theorem 2.2 decouple and the first equation reduces to the smooth anisotropic Allen–Cahn equation

$$\left\langle \frac{du}{dt}, v \right\rangle + \langle \mathcal{F}_\gamma(u), v\rangle + \big(\Psi'(u), v\big) = 0 \qquad \forall v \in H^1(\Omega), \qquad \text{a.e. in } (0,T).$$

Hence Theorem 2.2 provides the existence of solutions for strictly convex Kobayashi anisotropies if Ψ' is Lipschitz continuous. Unfortunately neither the logarithmic potential nor the obstacle potential satisfy the Lipschitz continuity condition.

A more general result by Elliott and Schätzle [52] considers a so-called fully anisotropic Allen–Cahn equation with obstacle potential. In contrast to the presented gradient flows a kinetic factor $\beta(\nabla u)$ is added in front of $\frac{du}{dt}$. As β is assumed to be 1-homogeneous the time derivative vanishes for $\nabla u = 0$. Due to this fact the authors do not apply the concept of weak solutions but show the existence of solutions in the viscosity sense (see [37]).

2.2.2 Cahn–Hilliard Equation

Regarding the Cahn–Hilliard equation we will mainly consider the isotropic case where $\gamma : \mathbb{R}^d \to \mathbb{R}$ takes the form

$$\gamma(x)^2 = \overline{\gamma}\|x\|^2$$

with some constant $\overline{\gamma} > 0$.

First we consider the logarithmic potential, i.e., $\theta > 0$. In this case $\widehat{\Psi}_\theta$ is differentiable on $(-1, 1)$ and hence Problem 2.2 can equivalently be written as the following variational equation.

Problem 2.4. *Find $u \in W(0,T;H^1(\Omega))$ and $w \in L^2(0,T;H^1(\Omega))$ such that $u(0) = u_0$ a.e. in Ω and*

$$\begin{aligned} \overline{\gamma}\,(\nabla u, \nabla v) - (w, v) + \left(\widehat{\psi}'_\theta(u), v\right) &= \theta_c\,(u, v) \qquad \forall v \in H^1(\Omega), \qquad \textit{a.e. in } (0,T), \\ \left\langle \frac{du}{dt}, v \right\rangle + (\nabla w, \nabla v) &= 0 \qquad\qquad\quad \forall v \in H^1(\Omega), \qquad \textit{a.e. in } (0,T). \end{aligned}$$

Existence of solutions for a vector-valued version of this problem was proved by Elliott and Luckhaus [51]. The summarized existence result for the binary case [35] reads as follows:

Theorem 2.3. *Let $u_0 \in \mathcal{K}$ and $|\,(u_0, 1)\,| < |\Omega|$. Then there is a unique solution u, w of Problem 2.4 that satisfies $u(t) \in \mathcal{K}$ a.e. in $(0,T)$. The following regularity results hold for u, w and $\frac{du}{dt}$:*

$$\begin{aligned} u &\in L^\infty(0,T;H^1(\Omega)) \cap L^2(0,T;H^2(\Omega)) \cap C(0,T;L^2(\Omega)), \\ \frac{du}{dt} &\in L^2(0,T;(H^1(\Omega))'), \\ (\sqrt{t})\frac{du}{dt} &\in L^2(0,T;H^1(\Omega)), \\ (\sqrt{t})w &\in L^\infty(0,T;H^1(\Omega)), \\ (\sqrt{t})\widehat{\Psi}'_\theta(u) &\in L^\infty(0,T;L^2(\Omega)). \end{aligned}$$

Furthermore, Elliott [50] established the following stability result for solutions of Problem 2.4.

Theorem 2.4. *Let $u_0 \in \mathcal{K}$ and $|\,(u_0, 1)\,| < |\Omega|$. Then for all $t \in [0,T]$ the solution u, w of Problem 2.4 satisfies*

$$E(u(t)) + \int_0^t (\nabla w(s), \nabla w(s))\; ds = E(u_0).$$

Note that the initial value is allowed to take the values -1 and 1 although $\widehat{\Psi}'_\theta(x)$ has infinite limits for $x \to \pm 1$. From the last regularity result and the limiting properties of Ψ_θ we also get that $|u|$ can only take the value 1 on subsets of $\Omega \times (0,T)$ with zero measure.

Since the potential $\widehat{\Psi}_\theta$ is no longer differentiable for $\theta = 0$ Problem 2.2 cannot be written as a variational equation is this case. However, the corresponding variational inequality can be simplified using the definition of $\mathcal{K}$.

Problem 2.5. *Find $u \in W(0,T,H^1(\Omega))$ and $w \in L^2(0,T,H^1(\Omega))$ such that $u(0) = u_0$ a.e. in Ω, $u(t) \in \mathcal{K}$ and*

$$\begin{aligned}\overline{\gamma}\,(\nabla u, \nabla(v-u)) - (w, v-u) \geq \theta_c\,(u, v-u) \qquad &\forall v \in \mathcal{K}, \qquad &&\text{a.e. in } (0,T),\\ \left\langle \frac{du}{dt}, v \right\rangle + (\nabla w, \nabla v) = 0 \qquad &\forall v \in H^1(\Omega), \qquad &&\text{a.e. in } (0,T).\end{aligned}$$

This isotropic Cahn–Hilliard equation with obstacle potential was analyzed by Blowey and Elliot [15]. A key result is the following existence and uniqueness theorem.

Theorem 2.5. *Let $u_0 \in \mathcal{K}$ and $|\,(u_0, 1)\,| < |\Omega|$. Then there is a unique solution u, w of Problem 2.5 that satisfies $u(t) \in \mathcal{K}$ a.e. in $(0,T)$ and*

$$u \in L^\infty(0,T;H^1(\Omega)) \cap H^1(0,T;H^1(\Omega)').$$

Furthermore, u satisfies $u \in C(\delta, T; H^1(\Omega))$ for all $\delta > 0$ and

$$E(u(t)) + \int_{t'}^{t} (\nabla w(s), \nabla w(s))\; ds \leq E(u(t'))$$

for all $t, t' \in [0,T]$ with $t' < t$. Hence E is a Lyapunov functional for Problem 2.5.

While anisotropic versions of the Cahn–Hiliard equation were discussed e.g. in Cahn and Taylor [30, 31], Rätz et al. [90] there are (to the author's knowledge) no comparable existence results for these equation.

3 Discretization of Cahn-Hilliard Equations with Logarithmic Potential

In this chapter we consider the discretization of the variational problems obtained from the gradient flows postulated in Chapter 2. We start by introducing notation for finite element discretizations and mass lumping for superposition operators. Then we give an overview of existing fully discrete approaches using the "method of lines". Finally we present a discretization using Rothe's method combined with spatial adaptivity.

3.1 Finite Element Spaces on Nonconforming Grids

Since we do not want to deal with boundary approximations we assume that the domain $\Omega \subset \mathbb{R}^d$ with $d = 1, 2, 3$ is a bounded, open, and nonempty polyhedron.

Definition 3.1. *A finite set $\mathcal{T} \subset 2^\Omega$ is a (simplicial) triangulation of Ω if each $\tau \in \mathcal{T}$ is a nonempty d-dimensional open simplex and*

$$\overline{\Omega} = \bigcup_{\tau \in \mathcal{T}} \overline{\tau}, \qquad \tau_1 \neq \tau_2 \;\Rightarrow\; \tau_1 \cap \tau_2 = \emptyset \qquad \forall \tau_1, \tau_2 \in \mathcal{T}.$$

A vertex/edge/2-face of some $\tau \in \mathcal{T}$ is called a node/edge/2-face of $\mathcal{T}$. The sets of all nodes and edges of $\mathcal{T}$ are denoted by $\mathcal{N}(\mathcal{T})$ and $\mathcal{E}(\mathcal{T})$, respectively. The diameter of $\tau \in \mathcal{T}$ is denoted by $h(\tau)$ and the maximal diameter of all elements is $h(\mathcal{T})$. The term face will be used for faces of arbitrary dimension and 2-dimensional faces are explicitly called 2-faces.

Definition 3.2. *A triangulation $\mathcal{T}$ of Ω is called conforming if for $\tau_1, \tau_2 \in \mathcal{T}$ with $\tau_1 \neq \tau_2$ the intersection $\overline{\tau_1} \cap \overline{\tau_2}$ is either empty, a vertex, an edge, or a 2-face of τ_1 and τ_2.*

We are especially interested in nonconforming triangulations obtained by adaptive refinement. While we do not want to restrict our considerations to conforming triangulations, we would still like to construct conforming piecewise polynomial finite element spaces.

Definition 3.3. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be triangulations of Ω. Then $\mathcal{T}_2$ is called a refinement of $\mathcal{T}_1$ if for all $\tau \in \mathcal{T}_1$ the set*

$$\{\tau' \in \mathcal{T}_2 : \tau' \cap \tau \neq \emptyset\}$$

is a triangulation of τ.

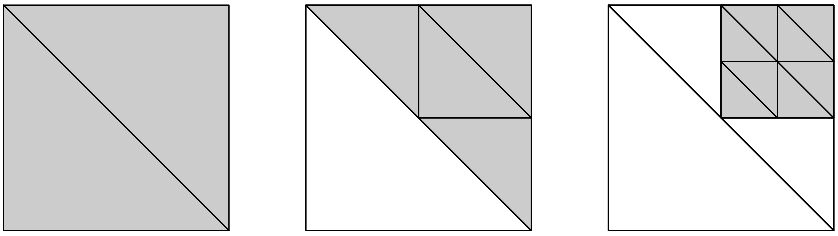


Figure 3.1: Level grids of a grid hierarchy with lower level elements dashed.

Definition 3.4. *$(\mathcal{T}_0, \dots, \mathcal{T}_j)$ is called a grid hierarchy on Ω if $\mathcal{T}_0$ is a conforming triangulation of Ω and if each $\mathcal{T}_i$, $i = 1, \dots, j$ is a conforming refinement of a subset of $\mathcal{T}_{i-1}$. $\mathcal{T}_i$ is called the i-th level grid of the grid hierarchy $(\mathcal{T}_0, \dots, \mathcal{T}_j)$.*

Strictly speaking we could call this a "conforming" grid hierarchy and drop the requirement of conforming level grids for a more general definition. However, we need the desired property for the following considerations and thus stick to this definition.

Since we allow refinements of real subsets of $\mathcal{T}_{i-1}$, the triangulations $\mathcal{T}_i$ with $i > 0$ do in general not cover the whole domain Ω as depicted in Figure 3.1. In case of a grid hierarchy the natural triangulation covering Ω to be used for computations is the so-called leaf triangulation or leaf grid.

Definition 3.5. *Let $(\mathcal{T}_0, \dots, \mathcal{T}_j)$ be a grid hierarchy on Ω. Then the leaf grid denoted by $\mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_j)$ is defined by*

$$\mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_j) = \mathcal{T}_j \cup \bigcup_{i=0}^{j-1} \{\tau \in \mathcal{T}_i : \tau \cap \tau' = \emptyset \ \forall \tau' \in \mathcal{T}_{i+1}\}.$$

Obviously the leaf grid of a grid hierarchy on Ω is itself a triangulation of Ω. As opposed to the level grids, it is in general not conforming. However, it is not as arbitrary as a general triangulation.

Lemma 3.1. *Let $(\mathcal{T}_0, \dots, \mathcal{T}_j)$ be a grid hierarchy. Then for $\tau_1, \tau_2 \in \mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_j)$ with $\tau_1 \neq \tau_2$ the intersection $\overline{\tau_1} \cap \overline{\tau_2}$ is either empty or a face of τ_1 or τ_2.*

Proof. For $\tau_1, \tau_2 \in \mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_j)$ with $\tau_1 \neq \tau_2$ we have $\tau_1 \in \mathcal{T}_{i_1}$ and $\tau_2 \in \mathcal{T}_{i_2}$. For $i_1 = i_2$ the assertion is clear because each level grid is a conforming triangulation. Without loss of generality we now assume $i_1 < i_2$. Then τ_2 is contained in one $\tau_2' \in \mathcal{T}_{i_1}$ and $\overline{\tau_1} \cap \overline{\tau_2'}$ is empty or a face of τ_2'. Hence $\overline{\tau_1} \cap \overline{\tau_2} = (\overline{\tau_1} \cap \overline{\tau_2'}) \cap \overline{\tau_2}$ must be empty or a face of τ_2. □

Figure 3.2 depicts the leaf grid of the grid hierarchy in Figure 3.1 on the left and two other nonconforming grids that do not have property shown in Lemma 3.1. If the intersection $\overline{\tau_1} \cap \overline{\tau_2}$ of to distinct elements $\tau_1, \tau_2 \in \mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_j)$ is only a face of τ_2 then there must be one vertex of τ_2 that is contained in the intersection but is not a vertex of τ_1.

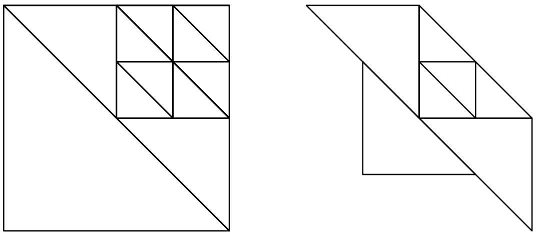


Figure 3.2: A leaf grid (left) and a non-leaf grid (right)

Definition 3.6. *Let $\mathcal{T}$ be a triangulation of Ω. Then a node $p \in \mathcal{N}(\mathcal{T})$ of $\mathcal{T}$ is called a hanging node if there is an element $\tau \in \mathcal{T}$ with $p \in \overline{\tau}$ but p is not a vertex of τ. The set of all hanging nodes of $\mathcal{T}$ is denoted by $\mathcal{H}(\mathcal{T})$.*

For finite element discretizations it is very common to only consider conforming triangulations of Ω. For such triangulations the k-th order Lagrangian finite element functions are the continuous functions on $\overline{\Omega}$ such that the restrictions to all elements $\tau \in \mathcal{T}$ are polynomials with degree of at most k. It is well known that these function spaces are conforming with respect to $H^1(\Omega)$, i.e., they are subspaces of $H^1(\Omega)$. However, the same definition does also lead to conforming spaces if it is used on a nonconforming triangulation.

For general nonconforming triangulations these spaces can degenerate in the sense that their dimension is small compared to a space on a conforming grid with a comparable number of elements and nodes. This is e.g. the case for the conforming space of piecewise linear functions on the right triangulation in Figure 3.2. In contrast to this we will see that conforming spaces on nonconforming leaf grids do in general not degenerate. We will restrict our considerations to the first-order case only.

Definition 3.7. *Let $\mathcal{T}$ be a triangulation of Ω. The first-order conforming finite element space is defined as*

$$\mathcal{S}(\mathcal{T}) := \{v \in C(\Omega) : v|_\tau \text{ is affine linear } \forall \tau \in \mathcal{T}\} \subset H^1(\Omega). \tag{3.1}$$

In case of a conforming triangulation a basis of $\mathcal{S}(\mathcal{T})$ is given by the well-known nodal basis functions. In order to deal with conforming finite element spaces on nonconforming grids we first introduce the nonconforming nodal basis functions.

Definition 3.8. *Let $\mathcal{T}$ be a triangulation of Ω. Then the nonconforming nodal basis function $\widehat{\lambda}_p \in L^2(\Omega)$ associated with $p \in \mathcal{N}(\mathcal{T})$ is defined as follows: For all $\tau \in \mathcal{T}$ there is an affine linear representative $\widehat{\lambda}_p|_\tau = \mu_{p,\tau} \in C(\overline{\tau})$ with $\mu_{p,\tau}(q) = \delta_{pq}$ for all vertices q of τ.*

For a conforming triangulation $\mathcal{T}$ this reduces to $\widehat{\lambda}_p \in \mathcal{S}(\mathcal{T})$ and

$$\widehat{\lambda}_p(q) = \delta_{pq} \qquad \forall p, q \in \mathcal{N}(\mathcal{T}),$$

i.e., the set $(\widehat{\lambda}_p)_{p\in\mathcal{N}(\mathcal{T})}$ is the conforming nodal basis of $\mathcal{S}(\mathcal{T})$. For a nonconforming triangulation $\mathcal{S}(\mathcal{T})$ is in general only a subspace of the nonconforming finite element space

$$\widehat{\mathcal{S}}(\mathcal{T}) := \operatorname{span}\{\widehat{\lambda}_p : p \in \mathcal{N}(\mathcal{T})\}.$$

However, in case of a leaf grid $\mathcal{T}$ it is possible to construct a basis of $\mathcal{S}(\mathcal{T})$ from the nonconforming nodal basis of $\widehat{\mathcal{S}}(\mathcal{T})$ that resembles the usual nodal basis functions where $\mathcal{T}$ is conforming.

Theorem 3.1. *Let $(\mathcal{T}_0, \dots, \mathcal{T}_j)$ be a grid hierarchy on Ω and $\mathcal{T} = \mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_j)$ the leaf grid. Then a basis of $\mathcal{S}(\mathcal{T})$ is given by*

$$\mathcal{B}(\mathcal{T}) := \left\{ \lambda_p = \widehat{\lambda}_p + \sum_{q\in\mathcal{H}(\mathcal{T})} a_{qp}\widehat{\lambda}_q : p \in \mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T}) \right\}.$$

Before proving Theorem 3.1 we show that hanging nodes can always be represented as linear combination of non-hanging nodes.

Lemma 3.2. *Let $(\mathcal{T}_0, \dots, \mathcal{T}_j)$ be a grid hierarchy on Ω and $\mathcal{T} = \mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_j)$ the leaf grid. Then for all $q \in \mathcal{H}(\mathcal{T})$ there are coefficients a_{qp} with $p \in \mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T})$ such that*

$$v(q) = \sum_{p\in\mathcal{N}(\mathcal{T})\setminus\mathcal{H}(\mathcal{T})} a_{qp}v(p) \qquad \forall v \in \mathcal{S}(\mathcal{T}).$$

Proof. (Lemma 3.2) The assertion trivially holds true for all $q \in \mathcal{H}(\mathcal{T}) \cap \mathcal{N}(\mathcal{T}_0) = \emptyset$. Now assume that it also holds true for all

$$q' \in \mathcal{H}_i := \mathcal{H}(\mathcal{T}) \cap \left(\bigcup_{k\le i} \mathcal{N}(\mathcal{T}_i) \right)$$

for some $0 \le i < j$ and let $q \in \mathcal{H}(\mathcal{T}) \cap \mathcal{N}(\mathcal{T}_{i+1})$. Then there is a $\tau \in \mathcal{T}$ such that $q \in \overline{\tau}$ but q is not a vertex of τ. Due to the definition of a grid hierarchy we must have $\tau \in \mathcal{T}_k$ for some $k \le i$. Then q can be written as convex combination of the vertices of τ, i.e.

$$q = \sum_{p\in\mathcal{N}(\mathcal{T}_k)} \bar{a}_{qp}p = \sum_{p\in\mathcal{N}(\mathcal{T})} \bar{a}_{qp}p$$

with $\bar{a}_{qp} = 0$ for all p that are not vertices of τ and, in particular, for $p \notin \mathcal{N}(\mathcal{T}_k)$. Since

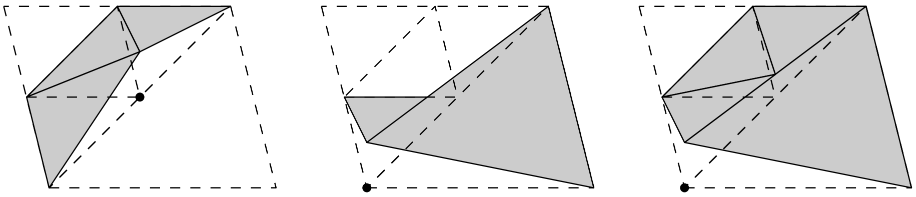


Figure 3.3: Nonconforming nodal basis functions for a hanging node (left) and a non-hanging node (middle) and a conforming nodal basis function for a non-hanging node (right).

all $v \in \mathcal{S}(\mathcal{T})$ are affine on $\overline{\tau}$ this implies

$$\begin{aligned}
v(q) &= \sum_{p\in\mathcal{N}(\mathcal{T}_k)\setminus\mathcal{H}(\mathcal{T})} \bar{a}_{qp} v(p) + \sum_{q'\in \underbrace{\mathcal{N}(\mathcal{T}_k)\cap\mathcal{H}(\mathcal{T})}_{\subset\mathcal{H}_k\subset\mathcal{H}_i}} \bar{a}_{qq'} v(q') \\
&= \sum_{p\in\mathcal{N}(\mathcal{T})\setminus\mathcal{H}(\mathcal{T})} \bar{a}_{qp} v(p) + \sum_{q'\in\mathcal{N}(\mathcal{T}_k)\cap\mathcal{H}(\mathcal{T})} \bar{a}_{qq'} \left(\sum_{p\in\mathcal{N}(\mathcal{T})\setminus\mathcal{H}(\mathcal{T})} a_{q'p} v(p) \right) \\
&= \sum_{p\in\mathcal{N}(\mathcal{T})\setminus\mathcal{H}(\mathcal{T})} \left(\bar{a}_{qp} + \sum_{q'\in\mathcal{N}(\mathcal{T}_k)\cap\mathcal{H}(\mathcal{T})} \bar{a}_{qq'} a_{q'p} \right) v(p).
\end{aligned}$$

We can define a_{qp} as the term in the parentheses. □

The coefficients in Lemma 3.2 can now be used to define conforming nodal basis functions for all non-hanging nodes. These need not be zero in the hanging nodes but take the proper value needed to ensure continuity. Figure 3.3 illustrates the difference between nonconforming and conforming nodal basis functions for the same node.

Proof. (Theorem 3.1) From Lemma 3.2 we get for all $v \in \mathcal{S}(\mathcal{T}) \subset \widehat{\mathcal{S}}(\mathcal{T})$

$$\begin{aligned}
v = \sum_{p\in\mathcal{N}(\mathcal{T})} \widehat{\lambda}_p v(p) &= \sum_{p\in\mathcal{N}(\mathcal{T})\setminus\mathcal{H}(\mathcal{T})} \widehat{\lambda}_p v(p) + \sum_{q\in\mathcal{H}(\mathcal{T})} \widehat{\lambda}_p \sum_{p\in\mathcal{N}(\mathcal{T})\setminus\mathcal{H}(\mathcal{T})} a_{qp} v(p) \\
&= \sum_{p\in\mathcal{N}(\mathcal{T})\setminus\mathcal{H}(\mathcal{T})} \lambda_p
\end{aligned}$$

and hence $\mathcal{S}(\mathcal{T}) \subset \operatorname{span}\mathcal{B}(\mathcal{T})$.

To see that also $\operatorname{span}\mathcal{B}(\mathcal{T}) \subset \mathcal{S}(\mathcal{T})$ we have to show that each λ_p with $p \in \mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T})$ is continuous. Since $q \in \mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T})$ is a vertex of all adjacent elements, all continuous representatives of restrictions of λ_p to these elements take the same value in q. Hence λ_p is continuous in q.

Now assume that $\lambda_p|_{\Omega_i}$ is continuous with

$$\Omega_i = \bigcup_{\tau \in \mathcal{T} \cap \bigcup_{k \le i} \mathcal{T}_k} \overline{\tau}$$

for some $0 \le i < j$ and let $\tau \in \mathcal{T} \cap \mathcal{T}_{i+1}$ and $q \in \overline{\tau}$ a vertex of τ. If $q \in \mathcal{H}(\mathcal{T})$ there is an element $\tau' \in \mathcal{T}$ such that $q \in \overline{\tau'}$ but q is not a vertex of τ'. Hence we have $\tau' \in \mathcal{T}_k$ for some $k \le i$ and thus $\tau' \in \Omega_i$.

If τ' is chosen as in the proof of Lemma 3.2 the value of $\lambda_p|_\tau$ in q is the interpolation of the values of $\lambda_p|_{\tau'}$ at the vertices of τ'. Thus $\lambda_p|_{\overline{\tau'} \cup \overline{\tau}}$ and (by continuity on Ω_i) also $\lambda_p|_{\Omega_i \cup \overline{\tau}}$ is continuous in q. Since this is true for all vertices of τ we have shown continuity of $\lambda_p|_{\Omega_i \cup \overline{\tau}}$ and hence of $\lambda_p|_{\Omega_{i+1}}$. Noting that $\lambda_p|_{\Omega_0}$ is continuous we have shown continuity on $\Omega_j = \mathcal{T}$ by induction. □

As we use multigrid solvers in the following chapters we now investigate the natural hierarchy of finite element spaces induced by a grid hierarchy. As a direct consequence of the definition of refinements we get nestedness of these spaces.

Lemma 3.3. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be triangulations of Ω such that $\mathcal{T}_2$ is a refinement of $\mathcal{T}_1$. Then $\widehat{\mathcal{S}}(\mathcal{T}_1)$ and $\mathcal{S}(\mathcal{T}_1)$ are subspaces of $\widehat{\mathcal{S}}(\mathcal{T}_2)$ and $\mathcal{S}(\mathcal{T}_2)$, respectively.*

□

The following lemma allows to use this property for grid hierarchies.

Lemma 3.4. *Let $(\mathcal{T}_0, \dots, \mathcal{T}_j)$ and $(\mathcal{T}'_0, \dots, \mathcal{T}'_{j'})$ be grid hierarchies with $j \le j'$, $\mathcal{T}_0 = \mathcal{T}'_0$, and $\mathcal{T}_i \subset \mathcal{T}'_i$ for all $0 < i \le j$. Then $\mathcal{L}(\mathcal{T}'_0, \dots, \mathcal{T}'_{j'})$ is a refinement of $\mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_j)$.*

Proof. Let $\tau \in \mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_{l-1})$. Then we have $\tau \in \mathcal{T}'_i$ for some $0 \le i \le j$. Hence

$$\{\tau' \in \mathcal{L}(\mathcal{T}'_0, \dots, \mathcal{T}'_{j'}) : \tau' \cap \tau \ne \emptyset\} = \mathcal{L}(\mathcal{T}'_i|_\tau, \dots, \mathcal{T}'_{j'}|_\tau)$$

is the leaf grid of the grid hierarchy $(\mathcal{T}'_i|_\tau, \dots, \mathcal{T}'_{j'}|_\tau)$ with

$$\mathcal{T}'_k|_\tau = \{\tau' \in \mathcal{T}'_k : \tau' \subset \tau\}$$

and hence itself a triangulation of τ. □

Corollary 3.1. *Let $(\mathcal{T}_0, \dots, \mathcal{T}_j)$ be a grid hierarchy and $\widetilde{\mathcal{T}}_l = \mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_l)$ for $l = 0, \dots, j$. Then $\widehat{\mathcal{S}}(\widetilde{\mathcal{T}}_{l-1})$ and $\mathcal{S}(\widetilde{\mathcal{T}}_{l-1})$ are subspaces of $\widehat{\mathcal{S}}(\widetilde{\mathcal{T}}_l)$ and $\mathcal{S}(\widetilde{\mathcal{T}}_l)$, respectively, for all $l = 1, \dots, j$.*

3.2 Mass Lumping and Superposition Operators

Although the bilinear forms $(\cdot, \cdot)$ and $(\nabla \cdot, \nabla \cdot)$ can be computed exactly for functions in $\mathcal{S}(\mathcal{T})$ it will be helpful to introduce an approximation of $(\cdot, \cdot)$ depending on $\mathcal{T}$.

Definition 3.9. *Let $\mathcal{T}$ be a triangulation of Ω that is the leaf grid of a grid hierarchy. Then the interpolation $I^{\mathcal{T}} : C(\overline{\Omega}) \to \mathcal{S}(\mathcal{T})$ is defined by*

$$I^{\mathcal{T}}(v) := \sum_{p \in \mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T})} v(p)\lambda_p$$

and the lumped L^2 inner product $(\cdot,\cdot)^{\mathcal{T}} : \mathcal{S}(\mathcal{T}) \times \mathcal{S}(\mathcal{T}) \to \mathbb{R}$ is defined by

$$\begin{aligned}(v,w)^{\mathcal{T}} &:= \int_\Omega I^{\mathcal{T}}(vw)(x)\,dx = \sum_{p \in \mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T})} v(p)w(p) \int_\Omega \lambda_p(x)\,dx && (3.2)\\ &\approx \int_\Omega v(x)w(x)\,dx = (v,w)\,. && (3.3)\end{aligned}$$

The lumped pseudo L^2 projection $P^{\mathcal{T}} : L^2(\Omega) \to \mathcal{S}(\mathcal{T})$ is defined by

$$\big(P^{\mathcal{T}} v, w\big)^{\mathcal{T}} = (v,w) \qquad \forall w \in \mathcal{S}(\mathcal{T}).$$

Since all λ_p with $p \in \mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T})$ are nonnegative and not identical to zero the weights in (3.2) are always positive. Hence the bilinear form $(\cdot,\cdot)^{\mathcal{T}}$ is an inner product. It is based on the approximation of the integral by a quadrature rule with the set of non-hanging nodes as quadrature points. In contrast to the exact inner product $(\cdot,\cdot)$ this bilinear form exhibits the discrete locality

$$\Big(\operatorname{supp}^{\mathcal{T}}(v) \cap \operatorname{supp}^{\mathcal{T}}(w) = \emptyset \quad \Rightarrow \quad (v,w)^{\mathcal{T}} = 0\Big) \qquad \forall v,w \in \mathcal{S}(\mathcal{T})$$

with respect to the discrete support $\operatorname{supp}(v)^{\mathcal{T}} := \{p \in \mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T}) : v(p) \neq 0\}$. This property is a discrete analogue of the continuous locality

$$\Big(\operatorname{supp}(v) \cap \operatorname{supp}(w) = \emptyset \quad \Rightarrow \quad (v,w) = 0\Big) \qquad \forall v,w \in L^2(\Omega).$$

This locality is important if superposition operators are considered. Remember that for a function $f : \mathbb{R} \to \mathbb{R}$ the associated superposition operator T_f on $L^p(\Omega)$ is given by $T_f v = f \circ v$. The above locality ensures that discretizations of superposition operators on $L^2(\Omega)$ are diagonal operators on $\mathcal{S}(\mathcal{T})$, which are the natural analogue of superposition operators on discrete spaces.

Remark 3.1. *Let $f : \mathbb{R} \to \mathbb{R}$ be continuous such that T with $Tv = f \circ v$ maps $L^2(\Omega) \to L^2(\Omega)$. Then T satisfies*

$$(Tv)|_U \text{ is independent of } (Tv)|_{\Omega\setminus U} \qquad \forall U \subset \Omega, U \text{ measurable.} \tag{3.4}$$

If we restrict T to $\mathcal{S}(\mathcal{T})$ it will not map $\mathcal{S}(\mathcal{T})$ to $\mathcal{S}(\mathcal{T})$. The latter can be ensured for a grid dependent $T^{\mathcal{T}} : \mathcal{S}(\mathcal{T}) \to \mathcal{S}(\mathcal{T})$ by using the weak definition

$$T^{\mathcal{T}} v \in \mathcal{S}(\mathcal{T}) : \qquad \big(T^{\mathcal{T}} v, w\big) = (Tv, w) \qquad \forall w \in \mathcal{S}(\mathcal{T}).$$

Unfortunately this operator does not satisfy (3.4)*. If we replace the inner product by the lumped L^2 product and define*

$$T^{\mathcal{T}} v \in \mathcal{S}(\mathcal{T}) : \qquad \left(T^{\mathcal{T}} v, w\right)^{\mathcal{T}} = (Tv, w)^{\mathcal{T}} \qquad \forall w \in \mathcal{S}(\mathcal{T})$$

the resulting operator does at least satisfy

$$(T^{\mathcal{T}} v)|_U \text{ is independent of } (T^{\mathcal{T}} v)|_{\Omega^{\mathcal{T}} \setminus U} \qquad \forall U \subset \Omega^{\mathcal{T}} := \mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T}).$$

This definition is equivalent to $T^{\mathcal{T}} = I^{\mathcal{T}} \circ (T|_{\mathcal{S}(\mathcal{T})})$.

3.3 Space–Time Discretizations

Now we summarize the present discretizations and corresponding convergence results for the Cahn–Hilliard equation. Again the present results only deal with the isotropic case where $\gamma : \mathbb{R}^d \to \mathbb{R}$ takes the form

$$\gamma(x)^2 = \overline{\gamma} \|x\|^2 \tag{3.5}$$

with some constant $\overline{\gamma} > 0$. All discretizations in this section use the "method of lines". This means that the time discretization is applied to the ordinary differential equation obtained by a fixed spatial discretization used for all time levels. Throughout this section we assume that $\mathcal{T}$ is a conforming triangulation. Furthermore, we consider a uniform time grid $0 = t_0 < t_1 < \dots$ with $t_k = k\Delta t$ for a constant time step size $\Delta t > 0$. Solutions corresponding to the k-th time level are indicated by a subscript k.

First we consider the logarithmic potential with $\theta > 0$, where the Cahn–Hilliard equation reduces to Problem 2.4. The discretization of this problem was discussed by Copetti and Elliott [35]. There, a fully implicit Euler scheme leading to the following sequence of discrete problems was proposed.

Problem 3.1. *Let $\Delta t > 0$ and $u_0^{\mathcal{T}} \in \mathcal{S}(\mathcal{T}) \cap \mathcal{K}$ some approximation of u_0. For $k = 1, \dots$ find $u_k^{\mathcal{T}}, w_k^{\mathcal{T}} \in \mathcal{S}(\mathcal{T})$ such that*

$$\begin{aligned} \overline{\gamma} \left(\nabla u_k^{\mathcal{T}}, \nabla v\right) - \theta_c \left(u_k^{\mathcal{T}}, v\right)^{\mathcal{T}} - \left(w_k^{\mathcal{T}}, v\right)^{\mathcal{T}} + \left(\widehat{\psi}_\theta'(u_k^{\mathcal{T}}), v\right)^{\mathcal{T}} &= 0 \qquad \forall v \in \mathcal{S}(\mathcal{T}),\\ \left(\frac{u_k^{\mathcal{T}} - u_{k-1}^{\mathcal{T}}}{\Delta t}, v\right)^{\mathcal{T}} + \left(\nabla w_k^{\mathcal{T}}, \nabla v\right) &= 0 \qquad \forall v \in \mathcal{S}(\mathcal{T}). \end{aligned}$$

Each discrete problem has the form of a saddle point problem

$$\begin{pmatrix} \overline{\gamma}A + F - M & -M \\ -M & -\Delta t A \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{w} \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}$$

for coefficient vectors $\underline{u}, \underline{w}$ and suitable right hand side vectors f, g. Here F is a monotone operator, M a symmetric positive definite matrix, and A a symmetric positive semidefinite matrix. Since $\overline{\gamma}A + F - M$ is not monotone this problem does in general not have a unique solution for all Δt. However, for small enough Δt the operator obtained by eliminating $\underline{w}$ is still monotone on the subspace where $\langle M\underline{u}, 1\rangle = 0$ or equivalently $(u, 1) = 0$ for the finite element function u represented by $\underline{u}$. This fact was used by Copetti and Elliott [35] to show the following existence result.

Theorem 3.2. *Let $\Delta t < 4\frac{\overline{\gamma}}{\theta_c^2}$, $u_0^{\mathcal{T}} \in \mathcal{S}(\mathcal{T}) \cap \mathcal{K}$ and $|\left(u_0^{\mathcal{T}}, 1\right)| < |\Omega|$. Then there is a unique solution $(u_k^{\mathcal{T}}, w_k^{\mathcal{T}}) \in \mathcal{S}(\mathcal{T})^2$ of Problem 3.1 that satisfy $\|u_k^{\mathcal{T}}\|_{L^\infty(\Omega)} < 1$ for all $k > 0$.*

Convergence was also established in [35] using the piecewise constant extensions $u^{\mathcal{T}}, w^{\mathcal{T}}$ of the discrete solution to the whole time interval given by

$$\begin{aligned} u^{\mathcal{T}} &\in L^2(0,T;H^1(\Omega)) : & u^{\mathcal{T}}(t) &= u_k^{\mathcal{T}} & \forall t &\in (t_{k-1}, t_k),\\ w^{\mathcal{T}} &\in L^2(0,T;H^1(\Omega)) : & w^{\mathcal{T}}(t) &= w_k^{\mathcal{T}} & \forall t &\in (t_{k-1}, t_k). \end{aligned}$$

Theorem 3.3. *Let $u_0 \in \mathcal{K}$ and $|(u_0, 1)| < |\Omega|$. Furthermore, let $u_0^{\mathcal{T}} = P^{\mathcal{T}} u_0$ and $\Delta t < 4\frac{\overline{\gamma}}{\theta_c^2}$. Then $u_0^{\mathcal{T}} \in \mathcal{K}$,$|\left(u_0^{\mathcal{T}}, 1\right)| < |\Omega|$ and the solution of Problem 3.1 converges to u in the sense that for all $\tau > 0$ its extension $u^{\mathcal{T}}$ satisfies $u^{\mathcal{T}} \to u$ in $L^2(\tau, T; L^2(\Omega))$ as $\Delta t, h(\mathcal{T}) \to 0$.*

Now we consider the obstacle potential ($\theta = 0$). An analogue time discretization for this case was introduced by Blowey and Elliot [16]. For simplicity the authors restrict their analysis to $\theta_c = 1$. As in the continuous case this leads to a variational inequality due to the non-differentiability.

Problem 3.2. *Let $\Delta t > 0$ and $u_0^{\mathcal{T}} \in \mathcal{S}(\mathcal{T}) \cap \mathcal{K}$ some approximation of u_0. For $k = 1, \dots$ find $u_k^{\mathcal{T}} \in \mathcal{S}(\mathcal{T}) \cap \mathcal{K}$ and $w_k^{\mathcal{T}} \in \mathcal{S}(\mathcal{T})$ such that*

$$\begin{aligned} \overline{\gamma}\left(\nabla u_k^{\mathcal{T}}, \nabla(v - u_k^{\mathcal{T}})\right) - \left(u_k^{\mathcal{T}}, v - u_k^{\mathcal{T}}\right)^{\mathcal{T}} - \left(w_k^{\mathcal{T}}, v - u_k^{\mathcal{T}}\right)^{\mathcal{T}} &\geq 0 \qquad \forall v \in \mathcal{S}(\mathcal{T}) \cap \mathcal{K},\\ \left(\frac{u_k^{\mathcal{T}} - u_{k-1}^{\mathcal{T}}}{\Delta t}, v\right)^{\mathcal{T}} + \left(\nabla w_k^{\mathcal{T}}, \nabla v\right) &= 0 \qquad \forall v \in \mathcal{S}(\mathcal{T}). \end{aligned}$$

Again each discrete problem takes the form of a saddle point problem

$$\begin{pmatrix} \overline{\gamma}A + \partial\chi_K - M & -M \\ -M & -\Delta t A \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{w} \end{pmatrix} \ni \begin{pmatrix} f \\ g \end{pmatrix}$$

where K is the representation of $\mathcal{S}(\mathcal{T}) \cap \mathcal{K}$ with respect to the nodal basis of $\mathcal{S}(\mathcal{T})$. As noted above the operator $\overline{\gamma}A + \partial\chi_K - M$ is in general not monotone and hence not the subdifferential of a convex function due to the “wrong” sign in front of the symmetric positive definite matrix M. This again leads to a time step restriction for the existence and uniqueness result in [16].

Theorem 3.4. *Let $u_0^{\mathcal{T}} \in \mathcal{S}(\mathcal{T}) \cap \mathcal{K}$ and $|\left(u_0^{\mathcal{T}}, 1\right)| < |\Omega|$. Then there is a solution $(u_k^{\mathcal{T}}, w_k^{\mathcal{T}}) \in \mathcal{S}(\mathcal{T})^2$ of Problem 3.2. For $\Delta t < 4\overline{\gamma}$ the order parameter $u_k^{\mathcal{T}}$ is unique. If, additionally, there is a node $p \in \mathcal{N}(\mathcal{T})$ with $|u_k^{\mathcal{T}}(p)| < 1$ then the chemical potential $w_k^{\mathcal{T}}$ is also unique.*

In order to overcome the time step restriction the authors also propose a semi-implicit time discretization where the concave part of the Ginzburg–Landau energy is discretized explicitly.

Problem 3.3. *Let $\Delta t > 0$ and $u_0^{\mathcal{T}} \in \mathcal{S}(\mathcal{T}) \cap \mathcal{K}$ some approximation of u_0. For $k = 1, \dots$ find $u_k^{\mathcal{T}} \in \mathcal{S}(\mathcal{T}) \cap \mathcal{K}$ and $w_k^{\mathcal{T}} \in \mathcal{S}(\mathcal{T})$ such that*

$$\begin{aligned} &\overline{\gamma}\left(\nabla u_k^{\mathcal{T}}, \nabla(v - u_k^{\mathcal{T}})\right) - \left(w_k^{\mathcal{T}}, v - u_k^{\mathcal{T}}\right)^{\mathcal{T}} \geq \left(u_{k-1}^{\mathcal{T}}, v - u_k^{\mathcal{T}}\right)^{\mathcal{T}} && \forall v \in \mathcal{S}(\mathcal{T}) \cap \mathcal{K}, \\ &\left(\frac{u_k^{\mathcal{T}} - u_{k-1}^{\mathcal{T}}}{\Delta t}, v\right)^{\mathcal{T}} + \left(\nabla w_k^{\mathcal{T}}, \nabla v\right) = 0 && \forall v \in \mathcal{S}(\mathcal{T}). \end{aligned}$$

Here the discrete saddle point problems take the form

$$\begin{pmatrix} \overline{\gamma}A + \partial\chi_K & -M \\ -M & -\Delta t A \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{w} \end{pmatrix} \ni \begin{pmatrix} f + M\underline{u}_{\text{old}} \\ g \end{pmatrix}$$

where the non-monotone operator $-M$ resulting from the concave part of Ginzburg–Landau energy is incorporated in the right hand side. The remaining operator $\overline{\gamma}A + \partial\chi_K$ is globally monotone and hence we get existence and uniqueness without time step restriction.

Theorem 3.5. *Let $u_0^{\mathcal{T}} \in \mathcal{S}(\mathcal{T}) \cap \mathcal{K}$ and $|\left(u_0^{\mathcal{T}}, 1\right)| < |\Omega|$. Then there is a solution $(u_k^{\mathcal{T}}, w_k^{T}) \in \mathcal{S}(\mathcal{T})^2$ of Problem 3.3 with unique $u_k^{\mathcal{T}}$ for all $\Delta t > 0$ and unique $w_k^{\mathcal{T}}$ if there is a node $p \in \mathcal{N}(\mathcal{T})$ with $|u_k^{\mathcal{T}}(p)| < 1$.*

The following error estimates by Blowey and Elliot [16] are again given in terms of piecewise constant extensions $u^{\mathcal{T}}, w^{\mathcal{T}} \in L^2(0, T; H^1(\Omega))$ as introduced above.

Theorem 3.6. *Let $u_0 \in \mathcal{K}$ and $|(u_0, 1)| < |\Omega|$. Furthermore, let $u_0^{\mathcal{T}} = P^{\mathcal{T}} u_0$ and $\Delta t < 2\overline{\gamma}$. Then the piecewise constant extension $(u^{\mathcal{T}}, w^{\mathcal{T}})$ of the solution to Problem 3.2 satisfies the error estimate*

$$\|u - u^{\mathcal{T}}\|^2_{L^\infty(0,T;(H^1(\Omega))')} + \|u - u^{\mathcal{T}}\|^2_{L^2(0,T;H^1(\Omega))} \leq C\left(\frac{h(\mathcal{T})^4}{\Delta t} + h(\mathcal{T})^2 + \Delta t\right)$$

with a constant C depending only on the solution (u, w) of Problem 2.5, the domain Ω and the shape regularity of $\mathcal{T}$.

While the existence result for the semi-implicit discretization in Problem 3.3 does not require any time step restriction, the following error estimate from [16] was only proved under a similar time step restriction as for the fully implicit case.

Theorem 3.7. *Let $u_0 \in \mathcal{K}$ and $|(u_0, 1)| < |\Omega|$. Furthermore, let $u_0^{\mathcal{T}} = P^{\mathcal{T}} u_0$ and $\Delta t < 2\overline{\gamma}$. Then the piecewise constant extension $(u^{\mathcal{T}}, w^{\mathcal{T}})$ of the solution to Problem 3.3 satisfies the error estimate*

$$\|u - u^{\mathcal{T}}\|^2_{L^\infty(0,T;(H^1(\Omega))')} + \|u - u^{\mathcal{T}}\|^2_{L^2(0,T;H^1(\Omega))} \leq C\left(\frac{h(\mathcal{T})^4}{\Delta t} + h(\mathcal{T})^2 + \Delta t\right)$$

with a constant C depending only on the solution (u, w) of Problem 2.5, the domain Ω and the shape regularity of $\mathcal{T}$.

Beside the above existence and convergence results there are many results for related equations. The Cahn–Hilliard equation with degenerate mobility incorporates a factor $b(u)$ with $b(-1) = b(1) = 0$ in front of ∇w in the second equation of Problem 2.4. For this equation well-posedness as well as convergence for one space dimension were established by Barrett et al. [8] for fully implicit and semi-implicit time discretization. Barrett et al. [9] analyzed a discretization for a similar equation with an additional potential equation. Error bounds in one and two space dimensions for a coupled Allen–Cahn/Cahn–Hilliard equation with logarithmic and obstacle potential were established by Barrett and Blowey [7].

Besides these results there are many results for Cahn–Hilliard equations with smooth double-well potentials (e.g. quartic polynomials). However, these results do in general rely heavily on the smoothness and do thus not carry over to the logarithmic potential and the obstacle potential considered here. Furthermore, many of the mentioned results have been generalized to vector-valued Cahn–Hilliard equations.

3.4 Rothe's Method

All discretizations presented so far follow the "method of lines" approach and use a fixed spatial discretization in each time step. In general, solutions to the presented phase field models are very smooth or even constant in large regions. These regions are occupied by a single pure or, in case of the logarithmic potential, almost pure phase. They are separated by a thin interface with a very sharp transition between different phases.

Due to this spatial variation of solutions it is reasonable to consider locally refined grids that allow for a local mesh size being only as small as needed to capture the local behavior. Since the interface moves through the domain in time, good locally refined grids will in general differ from one time step to the next. To allow for different grids in different time steps we will first discretize the evolution equation in time leading to a sequence of stationary continuous problems in appropriate Sobolev spaces. Each stationary problem is then discretized independently by finite elements. This approach is known as "Rothe's method" or "method of time slices" (see [44]). For the case of a linear parabolic partial differential equation it was analyzed extensively by Bornemann [18].

3.4.1 Semi-Implicit Time Discretization

In this subsection we consider a semi-implicit time discretization of the anisotropic Cahn–Hilliard equation with arbitrary $\theta \geq 0$ given by Problem 2.2. To this end let

$$0 = t_0 < t_1 < \dots$$

be a time grid with time steps $\Delta t_k = t_k - t_{k-1}$. Using an implicit Euler discretization for all terms except the gradient of the concave part of the potential in the Ginzburg–Landau energy leads to the following sequence of stationary problems:

Problem 3.4. *For $k = 1, \dots$ find $(u_k, w_k) \in H^1(\Omega) \times H^1(\Omega)$ such that*

$$\begin{aligned} \langle \mathcal{F}_\gamma(u_k), v - u_k\rangle - (w_k, v - u_k) + \widehat{\psi}_\theta(v) - \widehat{\psi}_\theta(u_k) &\geq \theta_c\,(u_{k-1}, v - u_k) && \forall v \in H^1(\Omega),\\ \left(\frac{u_k - u_{k-1}}{\Delta t_k}, v\right) + (\nabla w_k, \nabla v) &= 0 && \forall v \in H^1(\Omega). \end{aligned}$$

In will be shown that this time discretization has the advantage that no time step restriction needs to be imposed in order to guarantee the existence of solutions.

We will not analyze the convergence of the presented time discretization here but concentrate on the stationary problems for each time step t_k with $k > 0$ from now on. Since all these problems have the same structure we drop the index k for the time step and denote the solution from the previous time step by u_{old}. Then each stationary problem takes the form:

Problem 3.5. *Find $(u, w) \in H^1(\Omega) \times H^1(\Omega)$ such that*

$$\begin{aligned} \langle \mathcal{F}_\gamma(u), v - u\rangle + \widehat{\psi}_\theta(v) - \widehat{\psi}_\theta(u) - (w, v - u) &\geq \theta_c\,(u_{\text{old}}, v - u) && \forall v \in H^1(\Omega),\\ -\,(u, v) - \Delta t\,(\nabla w, \nabla v) &= -\,(u_{\text{old}}, v) && \forall v \in H^1(\Omega). \end{aligned}$$

In order to prove existence of solutions we will need continuity of the smooth nonquadratic functional

$$J_\gamma(v) = \int_\Omega \frac{1}{2}\gamma(\nabla v(x))^2 dx,$$

and lower semicontinuity of the nonsmooth nonlinearity $\widehat{\psi}_\theta$ in the Ginzburg–Landau energy. While continuity of the smooth nonlinearity is shown by elementary means in Lemma 2.5 we need a continuity result for superposition operators cited in the appendix to prove the following lemma.

Lemma 3.5. *For $\theta \geq 0$ the functional $\widehat{\psi}_\theta : H^1(\Omega) \to \mathbb{R} \cup \{\infty\}$ is lower semicontinuous. It is continuous on its domain $\operatorname{dom}\widehat{\psi}_\theta = \mathcal{K}$ which is a closed, convex, and nonempty set.*

Proof. A proof that $\mathcal{K}$ is closed and convex can be found in [72]. On $[-1, 1]$ the extended-valued function $\widehat{\Psi}_\theta$ coincides with the continuous function $\omega : \mathbb{R} \to \mathbb{R}$ given by

$$\omega(z) := \begin{cases} \widehat{\Psi}_\theta(z) & \text{if } z \in (-1, 1),\\ \theta \ln(2) & \text{else.} \end{cases}$$

By Corollary A.1 the superposition operator induced by ω is continuous from $L^2(\Omega)$ to $L^1(\Omega)$. Hence the functional

$$v \mapsto \int_\Omega \omega(v(x))dx$$

is in particular continuous on $\mathcal{K}$ equipped with the norm of $H^1(\Omega)$. Since this functional coincides with $\widehat{\psi}_\theta$ on $\mathcal{K}$ we find that $\widehat{\psi}_\theta$ is continuous on $\mathcal{K}$.

Now let $v^\nu \to v$ in $H^1(\Omega)$. If there is a ν_0 such that $v^\nu \notin \mathcal{K}$ for all $\nu \geq \nu_0$ we have

$$\liminf_{\nu\to\infty} \widehat{\psi}_\theta(v^\nu) = \infty \geq \widehat{\psi}_\theta(v).$$

If there is no such ν_0 there is a subsequence $v^{\nu_i} \in \mathcal{K}$ with $v^{\nu_i} \to v \in \mathcal{K}$. For any such subsequence we have

$$\lim_{i\to\infty} \widehat{\psi}_\theta(v^{\nu_i}) = \widehat{\psi}_\theta(v),$$

because $\widehat{\psi}_\theta$ is continuous on the closed set $\mathcal{K}$. Again we have shown

$$\liminf_{\nu\to\infty} \widehat{\psi}_\theta(v^\nu) \geq \widehat{\psi}_\theta(v).$$

□

A further ingredient in the proof of existence of solutions of Problem 3.5 is the coercivity of the convex energy J_γ associated with the nonlinear operator $\mathcal{F}_\gamma$. Unfortunately J_γ itself is not coercive on the whole space $H^1(\Omega)$ but only on the subspace where the variational equations in Problem 3.5 holds true. While this partial result would be sufficient for the proof of existence we show coercivity for a slightly modified equivalent problem in order to simplify the proof.

Testing the second equation in Problem 3.5 with $v = 1$ provides the mass conservation $(u, 1) = (u_{\text{old}}, 1)$. Hence we can add the term $\rho\,(u - u_{\text{old}}, 1)\,(1, v - u) = 0$ with $\rho > 0$ to the variational inequality without changing its solution. Defining

$$\left\langle \tilde{\mathcal{F}}_\gamma(v), \cdot \right\rangle := \langle \mathcal{F}_\gamma(v), \cdot \rangle + \rho\,(v, 1)\,(1, \cdot)\,, \qquad \tilde{f} := \theta_c u_{\text{old}} + \rho\,(u_{\text{old}}, 1)\,, \tag{3.6}$$

this leads to the equivalent semi-implicit problem:

Problem 3.6. *Find $(u, w) \in H^1(\Omega) \times H^1(\Omega)$ such that*

$$\begin{aligned} \left\langle \tilde{\mathcal{F}}_\gamma(u), v - u \right\rangle + \widehat{\psi}_\theta(v) - \widehat{\psi}_\theta(u) - (w, v - u) &\geq \left(\tilde{f}, v - u\right) && \forall v \in H^1(\Omega), \\ -\,(u, v) - \Delta t\,(\nabla w, \nabla v) &= -\,(u_{\text{old}}, v) && \forall v \in H^1(\Omega). \end{aligned}$$

Compared to Problem 3.5 this formulation has the advantage that the operator $\tilde{\mathcal{F}}_\gamma$ is strongly monotone with respect to the norm in $H^1(\Omega)$.

Lemma 3.6. *Assume that $\gamma : \mathbb{R}^d \to \mathbb{R}$ satisfies (A2). Then the gradients $\mathcal{F}_\gamma = \nabla J_\gamma$ and $\tilde{\mathcal{F}}_\gamma = \nabla \tilde{J}_\gamma$ are strongly monotone with respect to the semi-norm $|\cdot|_1$ and the norm $\|\cdot\|_1$, respectively, i.e.,*

$$\begin{aligned} \langle \mathcal{F}_\gamma(w) - \mathcal{F}_\gamma(v), w - v \rangle &\geq \underline{H}_{\gamma^2} |w - v|_1^2, \\ \left\langle \tilde{\mathcal{F}}_\gamma(w) - \tilde{\mathcal{F}}_\gamma(v), w - v \right\rangle &\geq \frac{\min\{\underline{H}_{\gamma^2}, \rho\}}{C_P} \|w - v\|_1^2. \end{aligned}$$

Proof. For strong monotonicity of $\mathcal{F}_\gamma$ we only need to apply the strong monotonicity of γ^2 from Lemma 2.4. The estimate for $\tilde{\mathcal{F}}_\gamma$ directly follows using the Poincaré inequality in Theorem A.2. □

Lemma 3.7. *Assume that $\gamma : \mathbb{R}^d \to \mathbb{R}$ satisfies (A2) and let $J_\gamma : H^1(\Omega) \to \mathbb{R}$ as defined in Lemma 2.5. Then the functional $\tilde{J}_\gamma : H^1(\Omega) \to \mathbb{R}$ given by*

$$\tilde{J}_\gamma(v) := J_\gamma(v) + \frac{\rho}{2}(v,1)^2$$

is strongly convex, continuous, and coercive. More precisely there is $C > 0$ such that

$$\tilde{J}_\gamma(v) \geq C\|v\|_1^2.$$

Proof. By Lemma 2.5 J_γ and thus $\tilde{J}_\gamma$ are continuous and convex. Strong convexity follows directly from strong convexity of γ^2 (see Lemma 2.4), strong convexity of the quadratic integral term, and the Poincaré inequality in Theorem A.2.

Similarly the coercivity of γ^2 (see Lemma 2.4) and the Poincaré inequality yield

$$\tilde{J}_\gamma(v) \geq \frac{\min\{\underline{H}_{\gamma^2}, \rho\}}{2}\left(|v|_1^2 + (v,1)^2\right) \geq \frac{\min\{\underline{H}_{\gamma^2}, \rho\}}{2C_P}\|v\|_1^2.$$

□

Theorem 3.8. *Assume that $\gamma : \mathbb{R}^d \to \mathbb{R}$ satisfies (A2) and that $|\,(u_{\text{old}}, 1)\,| < |\Omega|$. Then there is a solution $(u, w) \in \mathcal{K} \times H^1(\Omega)$ to Problem 3.6.*

Proof. We consider the Lagrange functional $L : \mathcal{K} \times H^1(\Omega) \to \mathbb{R}$,

$$L(u, w) = \tilde{J}_\gamma(u) + \widehat{\varphi}_\theta(u) - (\tilde{f}, u) - (u - u_{\text{old}}, w) - \frac{\Delta t}{2}|w|_1^2.$$

By Lemma 3.7 $\tilde{J}_\gamma$ is continuous and coercive. Together with Lemma 3.5 we find that $L(\cdot, w)$ is strictly convex, coercive, and continuous on the closed, convex, and nonempty set $\mathcal{K}$ for each $w \in H^1(\Omega)$. Furthermore, $L(u, \cdot)$ is continuous on $H^1(\Omega)$ for all $u \in \mathcal{K}$.

We can use Theorem A.4 to get existence if we can show

$$\lim_{\|w\|_1 \to \infty} \inf_{v \in \mathcal{K}} L(v, w) = -\infty. \tag{3.7}$$

To this end consider the decomposition

$$w = w_0 + w_c, \qquad w_c = \frac{(w,1)}{|\Omega|},$$

of $w \in H^1(\Omega)$. Then we have $(w_0, 1) = 0$ and thus by the Poincaré inequality (see Theorem A.2)

$$\begin{aligned} |\,(\operatorname{sgn} w_c - u_{\text{old}}, w_0)\,| = |\,(u_{\text{old}}, w_0)\,| &\leq \|u_{\text{old}}\|_0 \|w_0\|_0 \leq \sqrt{|\Omega|}\|w_0\|_1 \\ &\leq C_1\sqrt{|w_0|_1^2 + (w_0,1)^2} = C_1|w|_1. \end{aligned}$$

For w_c the identity $(|w_c|, 1) = |\,(w, 1)\,| = |w_c|\,|\Omega|$ and $|\,(u_{\text{old}}, 1)\,| < |\Omega|$ provide

$$\begin{aligned} -\left(\operatorname{sgn} w_c - u_{\text{old}}, w_c\right) &= -\left(|w_c|, 1\right) + \left(u_{\text{old}}, 1\right) w_c \\ &\leq -\left(|w_c|, 1\right) + |\,(u_{\text{old}}, 1)\,|\,|w_c| \\ &= -\underbrace{\left(1 - \frac{|\,(u_{\text{old}}, 1)\,|}{|\Omega|}\right)}_{=:C_2>0} |\,(w, 1)\,|. \end{aligned}$$

Hence with $a = |w|_1$ and $b = |\,(w, 1)\,|$ we get

$$-\left(\operatorname{sgn} w_c - u_{\text{old}}, w\right) - \frac{\Delta t}{2}|w|_1^2 \leq C_1 a - \frac{\Delta t}{2}a^2 - C_2 b.$$

To show that the right hand side tends to $-\infty$ uniformly for $\|w\|_1 \to \infty$ we note that the concave function $h : \mathbb{R}_0^+ \to \mathbb{R}$ with $h(a) = (C_1 + C_2)a - \frac{\Delta t}{2}a^2$ takes its maximum value at $a_0 = \frac{C_1+C_2}{\Delta t}$. This implies that

$$C_1 a + \frac{\Delta t}{2}a^2 = h(a) - C_2 a \leq h(a_0) - C_2 a = \underbrace{(C_1 + C_2)a_0 + \frac{\Delta t}{2}a_0^2}_{=:C_3>0} - C_2 a.$$

Together with the Poincaré inequality in Theorem A.2 we get

$$-\left(\operatorname{sgn} w_c - u_{\text{old}}, w\right) - \frac{\Delta t}{2}|w|_1^2 \leq C_3 - C_2(a + b) \leq C_3 - \frac{C_2}{\sqrt{C_P}}\|w\|_1.$$

Hence we have shown that

$$L(\operatorname{sgn} w_c, w) \leq \underbrace{\left(\max_{v\in\{-1,0,1\}} \tilde{J}_\gamma(v) + \widehat{\psi}_\theta(v) - (f, v) + C_3\right)}_{=\text{const}} - \frac{C_2}{\sqrt{C_P}}\|w\|_1$$

for all $w \in H^1(\Omega)$. Together with $\operatorname{sgn} w_c \in \mathcal{K}$ this provides (3.7). Now we can apply Theorem A.4 to obtain a saddle point $(u, w) \in \mathcal{K} \times H^1(\Omega)$ that satisfies

$$L(u, \mu) \leq L(u, w) \leq L(v, w) \qquad \forall (v, \mu) \in \mathcal{K} \times H^1(\Omega).$$

Standard arguments (see, e.g., [49, Chapter VI, Proposition 1.7]) lead to an optimality system for this saddle point problem given by a variational inequality for fixed w and a variational equation for fixed u. This system turns out to be Problem 3.6. □

Theorem 3.9. *Assume that $\gamma : \mathbb{R}^d \to \mathbb{R}$ satisfies (A2) and that $|\,(u_{\text{old}}, 1)\,| < |\Omega|$. Let $u, w \in H^1(\Omega)$ be a solution to Problem 3.6. Then u and ∇w are unique.*

Proof. Let $(u_1, w_1) \in \mathcal{K} \times H^1(\Omega)$ and $(u_2, w_2) \in \mathcal{K} \times H^1(\Omega)$ be two solutions to Problem 3.6. Testing the variational inequality for u_1 with u_2 and vice versa gives

$$\left\langle \tilde{\mathcal{F}}_\gamma(u_1) - \tilde{\mathcal{F}}_\gamma(u_2), u_1 - u_2 \right\rangle - (w_1 - w_2, u_1 - u_2) \leq 0,$$

while testing the variational equation for both solutions with $w_1 - w_2$ provides

$$-(u_1 - u_2, w_1 - w_2) = \Delta t |w_1 - w_2|_1^2.$$

Inserting this in the inequality and using the strong monotonicity of $\tilde{\mathcal{F}}_\gamma$ yields

$$C\|u_1 - u_2\|_1^2 + \Delta t |w_1 - w_2|_1^2 \leq \left\langle \tilde{\mathcal{F}}(u_1) - \tilde{\mathcal{F}}(u_2), u_1 - u_2 \right\rangle + \Delta t |w_1 - w_2|_1^2 \leq 0$$

and thus $u_1 = u_2$ and $\nabla w_1 = \nabla w_2$. □

Remark 3.2. *An alternative to adding the rank one term* $\rho\,(u,1)\,(1,\cdot)$ *in* (3.6) *would be to apply Theorem A.5 in the appendix to obtain another equivalent saddle point problem with a strongly monotone operator. While we have chosen the simpler approach with the rank one term here the other approach will be helpful for the fully implicit time discretization discussed in the next subsection.*

3.4.2 Fully Implicit Time Discretization

We can alternatively discretize the Cahn–Hilliard equation (Problem 2.2) by a fully implicit Euler scheme. This leads to a sequence of stationary problems which are similar to Problem 3.4:

Problem 3.7. *For* $k = 1, \dots$ *find* $(u_k, w_k) \in H^1(\Omega) \times H^1(\Omega)$ *such that*

$$\begin{aligned} \langle \mathcal{F}_\gamma(u_k), v - u_k\rangle - \langle \theta_c u_k, v - u_k\rangle - (w_k, v - u_k) + \widehat{\psi}_\theta(v) - \widehat{\psi}_\theta(u_k) \geq 0 \quad &\forall v \in H^1(\Omega), \\ \left(\frac{u_k - u_{k-1}}{\Delta t_k}, v\right) + (\nabla w_k, \nabla v) = 0 \quad &\forall v \in H^1(\Omega). \end{aligned}$$

Again we concentrate on a single spatial problem and drop the index k. Then each stationary problem for the fully implicit time discretization takes the form:

Problem 3.8. *Find* $(u, w) \in H^1(\Omega) \times H^1(\Omega)$ *such that*

$$\begin{aligned} \langle \mathcal{F}_\gamma(u), v - u\rangle - \theta_c\,(u, v - u) + \widehat{\psi}_\theta(v) - \widehat{\psi}_\theta(u) - (w, v - u) \geq 0 \qquad &\forall v \in H^1(\Omega), \\ -(u, v) - \Delta t\,(\nabla w, \nabla v) = -(u_{\text{old}}, v) \quad &\forall v \in H^1(\Omega). \end{aligned}$$

By the same arguments as for the semi-implicit discretization we have mass conservation and adding the term $\rho\,(u - u_{\text{old}}, 1)\,(1, v - u) = 0$ to the variational inequality leads to the equivalent

Problem 3.9. *Find* $(u, w) \in H^1(\Omega) \times H^1(\Omega)$ *such that*

$$\begin{aligned} \langle \tilde{\mathcal{F}}_\gamma(u), v - u\rangle - \theta_c\,(u, v - u) + \widehat{\psi}_\theta(v) - \widehat{\psi}_\theta(u) - (w, v - u) \geq (\tilde{f}, v - u) \quad &\forall v \in H^1(\Omega), \\ -(u, v) - \Delta t\,(\nabla w, \nabla v) = -(u_{\text{old}}, v) \quad &\forall v \in H^1(\Omega). \end{aligned}$$

Here $\tilde{\mathcal{F}}_\gamma$ is defined as in the previous subsection but the right hand side is now

$$\tilde{f} = \rho\left(u_{\text{old}}, 1\right).$$

In contrast to the semi-implicit scheme this modification does not guarantee monotonicity of $\tilde{\mathcal{F}}_\gamma - \theta_c\left(\cdot, \cdot\right)$. The reason for this is the "wrong" sign in front of $\theta_c\left(u, v-u\right)$ resulting from the concave part in the Ginzburg–Landau energy. We noted in the previous section (see Theorem 3.2) for a fully discrete scheme that the isotropic analogue of the operator $\mathcal{F}_\gamma - \theta_c\left(\cdot, \cdot\right)$ is in general only monotone on the subspace where the variational equation holds and if the time step restriction $\Delta t < 4\overline{\gamma}/\theta_c^2$ is satisfied. A corresponding result could be shown here. However, having monotonicity only on this subspace will in general prohibit algorithms where the linear equation only holds in the limiting case. Hence it is desirable to have a globally monotone operator associated with a globally convex functional instead.

Having this in mind we consider a different modification of Problem 3.8 using Theorems A.5 and A.6 in the appendix, which allows to compensate a certain fraction of a negative semidefinite operator. More precisely we can apply Theorem A.5 with $\alpha = \frac{\theta_c+\epsilon}{2}$ for any $\epsilon \geq 0$ to obtain:

Problem 3.10. *Find $(u, \tilde{w}) \in H^1(\Omega) \times H^1(\Omega)$ such that*

$$\begin{aligned} \left\langle \mathcal{F}_\gamma^\epsilon(u), v-u \right\rangle + \widehat{\psi}_\theta(v) - \widehat{\psi}_\theta(u) - b^\epsilon(\tilde{w}, v-u) &\geq \left(f^\epsilon, v-u\right) \qquad \forall v \in H^1(\Omega), \\ -b^\epsilon(u,v) - \Delta t\left(\nabla\tilde{w}, \nabla v\right) &= -\left(u_{\text{old}}, v\right) \qquad \forall v \in H^1(\Omega). \end{aligned}$$

Here we use the modified operator, bilinear form, and right hand side

$$\begin{aligned} \mathcal{F}_\gamma^\epsilon(v) &:= \mathcal{F}_\gamma(v) - \Delta t\frac{(\theta_c+\epsilon)^2}{4}\left(\nabla v, \nabla\cdot\right) + \epsilon\left(v, \cdot\right), \\ b^\epsilon(v, \cdot) &:= \left(v, \cdot\right) - \Delta t\frac{\theta_c+\epsilon}{2}\left(\nabla v, \nabla\cdot\right), \\ f^\epsilon &:= \frac{\theta_c+\epsilon}{2}u_{\text{old}}. \end{aligned}$$

Theorem 3.10. *For any $\epsilon \geq 0$ Problem 3.10 is equivalent to Problem 3.8 in the sense that $(u, \tilde{w}) \in H^1(\Omega) \times H^1(\Omega)$ is a solution to Problem 3.10 if and only if $(u, w) \in H^1(\Omega) \times H^1(\Omega)$ with $w = \tilde{w} - \frac{\theta_c+\epsilon}{2}u$ is a solution of Problem 3.8.*

Proof. Apply Theorem A.5 with $\alpha = \frac{\theta_c+\epsilon}{2}$. □

Theorem 3.11. *Assume that $\gamma : \mathbb{R}^d \to \mathbb{R}$ satisfies (A2), and that $|\left(u_{\text{old}}, 1\right)| < |\Omega|$ and $\Delta t < \underline{H}_{\gamma^2}\frac{4}{\theta_c^2}$ hold true. Then there is a solution $(u, \tilde{w}) \in \mathcal{K} \times H^1(\Omega)$ to Problem 3.10 for all $\epsilon \geq 0$.*

Proof. We show the existence of a solution to Problem 3.8 for a suitable choice of $\epsilon \geq 0$. This is sufficient since Problem 3.8 is equivalent to Problem 3.10 for all $\epsilon \geq 0$. Hence, for the rest of the proof we select a fixed $\epsilon > 0$ such that the first inequality in

$$\Delta t < \underline{H}_{\gamma^2}\frac{4}{(\theta_c+\epsilon)^2} < \underline{H}_{\gamma^2}\frac{4}{\theta_c^2}$$

holds true. This is possible due to the strict inequality in the time step restriction. Then, by the same arguments used in the proof of Lemma 3.7, the functional $J_\gamma^\epsilon : H^1(\Omega) \to \mathbb{R}$ defined by

$$J_\gamma^\epsilon(v) := J_\gamma(v) - \frac{\Delta t(\theta_c + \epsilon)^2}{8}|v|_1^2 + \frac{\epsilon}{2}\|v\|_0^2$$

is strongly convex, continuous, and coercive with the constant

$$\frac{1}{2}\min\left\{\underline{H}_{\gamma^2} - \Delta t\frac{(\theta_c+\epsilon)^2}{4}, \epsilon\right\} > 0.$$

Note that the concave part $-|v|_1^2$ is dominated by the convex part J_γ due to the special selection of ϵ.

Now we can essentially proceed as in the proof of Theorem 3.8 and show existence of a saddle point of the Lagrange functional $\tilde{L} : \mathcal{K} \times H^1(\Omega) \to \mathbb{R}$,

$$\tilde{L}(u, \tilde{w}) := J_\gamma^\epsilon(u) + \widehat{\varphi}_\theta(u) - (f^\epsilon, u) - b^\epsilon(u - u_{\text{old}})(\tilde{w}) - \frac{\Delta t}{2}|\tilde{w}|_1^2.$$

Again we have to show

$$\lim_{\|\tilde{w}\|_{H^1(\Omega)}\to\infty} \inf_{v\in\mathcal{K}} \tilde{L}(v, \tilde{w}) = -\infty$$

in order to apply Theorem A.4. To this end we note that by Lemma A.2 we have

$$\tilde{L}(u, \tilde{w}) = L(u, w) := J_\gamma(u) - \frac{\theta_c}{2}\|u\|_0^2 + \widehat{\varphi}_\theta(u) - (u - u_{\text{old}}, w) - \frac{\Delta t}{2}|w|_1^2$$

with $w = \tilde{w} - \frac{\theta_c+\epsilon}{2}u$. By the same arguments as in the proof of Theorem 3.8 we get

$$\tilde{L}(\operatorname{sgn} w_c, \tilde{w}) = L(\operatorname{sgn} w_c, w) \le C - C\|w\|_1$$

for a constant $C > 0$. Together with $\|\operatorname{sgn} w_c\|_1 = \|\operatorname{sgn} w_c\|_0 \le \sqrt{|\Omega|}$ and the inverse triangle inequality this implies

$$\begin{aligned}\tilde{L}(\operatorname{sgn} w_c, \tilde{w}) &\le C - C\left\|\tilde{w} - \frac{\theta_c+\epsilon}{2}\operatorname{sgn} w_c\right\|_1 \\ &\le C - C\left(\|\tilde{w}\|_1 - \frac{\theta_c+\epsilon}{2}\|\operatorname{sgn} w_c\|_1\right) \\ &\le C - C\|\tilde{w}\|_1 + C\sqrt{\Omega}\frac{\theta_c+\epsilon}{2}.\end{aligned}$$

Now Theorem A.4 provides the existence of a saddle point of $\tilde{L}$ that is a solution to Problem 3.10. □

Theorem 3.12. *Assume that $\gamma : \mathbb{R}^d \to \mathbb{R}$ satisfies (A2), and that $|\,(u_{\text{old}}, 1)\,| < |\Omega|$ and $\Delta t < \underline{H}_{\gamma^2}\frac{4}{\theta_c^2}$ hold true. Let $(u, \tilde{w}) \in H^1(\Omega) \times H^1(\Omega)$ be a solution to Problem 3.10 for some $\epsilon \ge 0$. Then u and $\nabla\tilde{w}$ are unique.*

Proof. By the same arguments as in Lemma 3.6 we get

$$\left\langle \mathcal{F}^{\epsilon}_{\gamma}(w) - \mathcal{F}^{\epsilon}_{\gamma}(v), w - v \right\rangle \geq \min \left\{ \underline{H}_{\gamma^2} - \Delta t \frac{(\theta_c + \epsilon)^2}{4}, \epsilon \right\} \|w - v\|_1^2.$$

Using this strong monotonicity of the operator $\mathcal{F}^{\epsilon}_{\gamma}$ we can show uniqueness as in the proof of Theorem 3.9. □

3.4.3 Spatial Discretization

Now we consider the spatial discretization of the stationary problems obtained at each time step t_k with $k > 0$ by semi- or fully-implicit time discretization. We will again drop the index k for the time step and denote the discrete solution from the previous time step as u_{old}.

As noted before we will consider locally refined grids to accommodate the strong spatial variations of solutions. In order to obtain grids that allow for nontrivial conforming finite element spaces we will construct a grid hierarchy starting with a conforming triangulation $\mathcal{T}_0$ of Ω on the 0-th level. Successive local refinement then leads to a grid hierarchy $(\mathcal{T}_0, ..., \mathcal{T}_j)$. During each refinement cycle computations are carried out on the current leaf grid $\mathcal{T} = \mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_j)$. We consider first-order conforming finite elements only. For the semi-implicit time discretization in Problem 3.6 this leads to discrete problems of the form:

Problem 3.11. *Find $(u^{\mathcal{T}}, w^{\mathcal{T}}) \in \mathcal{S}(\mathcal{T}) \times \mathcal{S}(\mathcal{T})$ such that*

$$\begin{aligned} \left\langle \tilde{\mathcal{F}}_{\gamma}(u^{\mathcal{T}}), v - u^{\mathcal{T}} \right\rangle + \left(\widehat{\Psi}_{\theta}(v) - \widehat{\Psi}_{\theta}(u^{\mathcal{T}}), 1\right)^A & \\ - \left(w^{\mathcal{T}}, v - u^{\mathcal{T}}\right)^B \geq \left(\tilde{f}, v - u^{\mathcal{T}}\right)^C & \qquad \forall v \in \mathcal{S}(\mathcal{T}), \quad (3.8) \\ - \left(u^{\mathcal{T}}, v\right)^D - \Delta t \left(\nabla w^{\mathcal{T}}, \nabla v\right) = - \left(u_{\text{old}}, v\right)^E & \qquad \forall v \in \mathcal{S}(\mathcal{T}). \quad (3.9) \end{aligned}$$

Here $(\cdot,\cdot)^A, \dots, (\cdot,\cdot)^E$ are approximations to the L^2 inner product to be selected carefully. If the solution is represented in terms of the conforming nodal basis selecting $(\cdot,\cdot)^A = (\cdot,\cdot)$ would couple coefficients from neighboring nodes within the nonsmooth nonlinearity, leading to hard to solve algebraic problems. As a remedy we choose the lumped L^2 inner product $(\cdot,\cdot)^A = (\cdot,\cdot)^{\mathcal{T}}$ that (as discussed in Section 3.2) allows to inherit the locality of the superposition operator. This discrete locality is equivalent to the fact that the nonsmooth nonlinearity decouples with respect to the coefficients.

In order to preserve the symmetry necessary for the saddle point structure, $(\cdot,\cdot)^B$ and $(\cdot,\cdot)^D$ should be the same. Since $\tilde{f} = \theta_c u_{\text{old}}$+const it is reasonable to use $(\cdot,\cdot)^C = (\cdot,\cdot)^E$ in order to approximate both appearances of u_{old} using the same inner product. It remains to select $(\cdot,\cdot)^D$ and $(\cdot,\cdot)^E$. Since both inner products result from the time discretization it is reasonable to select these products the same. The argument for this is that for $\Delta t \to 0$ the variational equation (3.9) degenerates to

$$\left(u^{\mathcal{T}}, v\right)^D = \left(u_{\text{old}}, v\right)^E \qquad v \in \mathcal{S}(\mathcal{T}).$$

If the inner products are not the same the induced operator $u_{\text{old}} \mapsto u^{\mathcal{T}}$ is in general not a projection which will introduce errors for $\Delta t \to 0$.

In order to enforce the mass conservation

$$\left(u^{\mathcal{T}}, 1\right)^D = (u_0, 1) = (u_{\text{old}}, 1)^D$$

the product should guarantee $(v, 1)^D = (v, 1)$ for $v = u_{\text{old}}$ and for each $v \in \mathcal{S}(\mathcal{T})$. In view of the fact that u_{old} is a finite element function on a different grid the lumped L^2 inner product will in general not satisfy this since it does only integrate piecewise linear functions on the current grid exactly. The same is true if the lumped L^2 inner product with respect to the grid from the last time step is selected. In principle this is possible if the lumped L^2 inner product with respect to a finite element space that contains u_{old} and $\mathcal{S}(\mathcal{T})$ is selected. However, such a product does no longer have the locality property of $(\cdot,\cdot)^{\mathcal{T}}$ on $\mathcal{S}(\mathcal{T})$ and seems to be quite arbitrary. Hence it does not give any benefit over using $(\cdot,\cdot)$ and we select the latter instead. Inserting all products we obtain the following discretization of Problem 3.6:

Problem 3.12. *Find* $(u^{\mathcal{T}}, w^{\mathcal{T}}) \in \mathcal{S}(\mathcal{T}) \times \mathcal{S}(\mathcal{T})$ *such that*

$$\begin{aligned} \left\langle \tilde{\mathcal{F}}_\gamma(u^{\mathcal{T}}), v - u^{\mathcal{T}} \right\rangle + \widehat{\psi}_\theta^{\mathcal{T}}(v) - \widehat{\psi}_\theta^{\mathcal{T}}(u^{\mathcal{T}}) - \left(w^{\mathcal{T}}, v - u^{\mathcal{T}}\right) &\geq \left(\tilde{f}, v - u^{\mathcal{T}}\right) && \forall v \in \mathcal{S}(\mathcal{T}), \\ -\left(u^{\mathcal{T}}, v\right) - \Delta t \left(\nabla w^{\mathcal{T}}, \nabla v\right) &= -\left(u_{\text{old}}, v\right) && \forall v \in \mathcal{S}(\mathcal{T}). \end{aligned}$$

Analogously we get the following discretization of Problem 3.10:

Problem 3.13. *Find* $(u^{\mathcal{T}}, \tilde{w}^{\mathcal{T}}) \in \mathcal{S}(\mathcal{T}) \times \mathcal{S}(\mathcal{T})$ *such that*

$$\begin{aligned} \left\langle \mathcal{F}_\gamma^\epsilon(u^{\mathcal{T}}), v - u^{\mathcal{T}} \right\rangle + \widehat{\psi}_\theta^{\mathcal{T}}(v) - \widehat{\psi}_\theta^{\mathcal{T}}(u^{\mathcal{T}}) - b^\epsilon(\tilde{w}^{\mathcal{T}}, v - u^{\mathcal{T}}) &\geq \left(f^\epsilon, v - u^{\mathcal{T}}\right) && \forall v \in \mathcal{S}(\mathcal{T}), \\ -b^\epsilon(u^{\mathcal{T}}, v) - \Delta t \left(\nabla \tilde{w}^{\mathcal{T}}, \nabla v\right) &= -\left(u_{\text{old}}, v\right) && \forall v \in \mathcal{S}(\mathcal{T}). \end{aligned}$$

In both cases we used the discrete approximation

$$\widehat{\psi}_\theta^{\mathcal{T}}(v) = \left(\widehat{\Psi}_\theta(v), 1\right)^{\mathcal{T}} = \int_\Omega I^{\mathcal{T}}\left(\widehat{\Psi}_\theta(v)\right)(x)\,dx = \sum_{p \in \mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T})} \widehat{\Psi}_\theta(v(p)) \int_\Omega \lambda_p(x)\,dx$$

of $\widehat{\psi}_\theta(v)$ for $v \in \mathcal{S}(\mathcal{T})$, where $I^{\mathcal{T}}$ is the interpolation operator introduced in Definition 3.9. Since the existence and uniqueness results and their proofs are essentially the same for both discretizations we will state and prove them in parallel.

Existence of solutions can be shown using the same arguments as for the continuous problem. We only have to show lower semicontinuity of the discrete nonlinearity. In contrast to the continuous case this can be done by elementary arguments.

Lemma 3.8. *For* $\theta \geq 0$ *the functional* $\widehat{\psi}_\theta^{\mathcal{T}} : \mathcal{S}(\mathcal{T}) \to \mathbb{R} \cup \{\infty\}$ *is lower semicontinuous. It is continuous on its domain* $\operatorname{dom} \widehat{\psi}_\theta^{\mathcal{T}} = \mathcal{K} \cap \mathcal{S}(\mathcal{T})$ *which is a closed, convex, and nonempty set.*

Proof. The domain of $\widehat{\psi}_\theta^\mathcal{T}$ is given by the closed convex set

$$\mathcal{K} \cap \mathcal{S}(\mathcal{T}) = \big\{ v \in \mathcal{S}(\mathcal{T}) : |v(p)| \leq 1 \, \forall p \in \mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T}) \big\}.$$

Hence on $\mathcal{K} \cap \mathcal{S}(\mathcal{T})$ the functional $\widehat{\psi}_\theta^\mathcal{T}$ is the sum of continuous functions and thus itself continuous on this set. Lower semicontinuity on the whole set $\mathcal{S}(\mathcal{T})$ can now be shown analogously to Lemma 3.5. □

Theorem 3.13. *Assume that $\gamma : \mathbb{R}^d \to \mathbb{R}$ satisfies (A2), and that $|\,(u_{\text{old}}, 1)\,| < |\Omega|$ holds. Then there is a solution $(u^\mathcal{T}, w^\mathcal{T}) \in \mathcal{K} \cap \mathcal{S}(\mathcal{T}) \times \mathcal{S}(\mathcal{T})$ to Problem 3.12, and $u^\mathcal{T}$ and $\nabla w^\mathcal{T}$ are unique.*

Proof. The proofs for existence and uniqueness are the same as for Theorem 3.8 and Theorem 3.9 with $H^1(\Omega)$, $\mathcal{K}$, and $\widehat{\psi}_\theta$ replaced by $\mathcal{S}(\mathcal{T})$, $\mathcal{K} \cap \mathcal{S}(\mathcal{T})$, and $\widehat{\psi}_\theta^\mathcal{T}$, respectively. □

Theorem 3.14. *Assume that $\gamma : \mathbb{R}^d \to \mathbb{R}$ satisfies (A2), and that $|\,(u_{\text{old}}, 1)\,| < |\Omega|$ and $\Delta t < \underline{H}_{\gamma^2} \frac{4}{\theta_c^2}$ hold true. Then there is a solution $(u^\mathcal{T}, \tilde{w}^\mathcal{T}) \in \mathcal{K} \cap \mathcal{S}(\mathcal{T}) \times \mathcal{S}(\mathcal{T})$ to Problem 3.13 and $u^\mathcal{T}$ and $\nabla \tilde{w}^\mathcal{T}$ are unique.*

Proof. The proofs for existence and uniqueness are the same as for Theorem 3.11 and Theorem 3.12 with $H^1(\Omega)$, $\mathcal{K}$, and $\widehat{\psi}_\theta$ replaced by $\mathcal{S}(\mathcal{T})$, $\mathcal{K} \cap \mathcal{S}(\mathcal{T})$, and $\widehat{\psi}_\theta^\mathcal{T}$, respectively. □

Lemma 3.9. *Let $\theta > 0$. Then the solutions $(u^\mathcal{T}, w^\mathcal{T})$ of Problem 3.12 and $(u^\mathcal{T}, \tilde{w}^\mathcal{T})$ of Problem 3.13 satisfy $\|u\|_\infty < 1$.*

Proof. From $\widehat{\Psi}_\theta'(t) \to \pm\infty$ for $t \to \pm 1$ we get $\partial \widehat{\Psi}_\theta(-1) = \partial \widehat{\Psi}_\theta(1) = \emptyset$. Hence by Theorem A.1 we have $\partial \widehat{\psi}_\theta^\mathcal{T}(v) = \emptyset$ for all $v \in \mathcal{S}(\mathcal{T})$ with $v(x) = 1$ for some $x \in \Omega$.

Noting that Problem 3.12 and Problem 3.13 are equivalent to operator inclusions incorporating $\partial \widehat{\psi}_\theta^\mathcal{T}$ we find that there cannot be any $x \in \Omega$ with $|u^\mathcal{T}(x)| = 1$. The assertion then follows from continuity of $u^\mathcal{T}$. □

Since $\widehat{\Psi}_\theta$ is differentiable on $(-1, 1)$, we can now rewrite Problem 3.12 and Problem 3.13 as variational equations for $\theta > 0$. This is true because the variational inequality is equivalent to a minimization problem for fixed $w^\mathcal{T}$ that is itself equivalent to a variational equation if we exploit differentiability of the energy (apply Proposition 1.2 in [49, Chapter II] twice with different splittings of the functional).

Problem 3.14. *Find $(u^\mathcal{T}, w^\mathcal{T}) \in \mathcal{S}(\mathcal{T}) \times \mathcal{S}(\mathcal{T})$ such that*

$$\begin{aligned} \Big\langle \tilde{\mathcal{F}}_\gamma(u^\mathcal{T}), v \Big\rangle + \Big\langle \nabla \widehat{\psi}_\theta^\mathcal{T}(u^\mathcal{T}), v \Big\rangle - \left(w^\mathcal{T}, v\right) &= \Big(\tilde{f}, v\Big) \qquad && \forall v \in \mathcal{S}(\mathcal{T}), \\ -\left(u^\mathcal{T}, v\right) - \Delta t \left(\nabla w^\mathcal{T}, \nabla v\right) &= -\left(u_{\text{old}}, v\right) && \forall v \in \mathcal{S}(\mathcal{T}). \end{aligned}$$

Problem 3.15. *Find $(u^{\mathcal{T}}, \tilde{w}^{\mathcal{T}}) \in \mathcal{S}(\mathcal{T}) \times \mathcal{S}(\mathcal{T})$ such that*

$$\begin{aligned}\langle \mathcal{F}^{\epsilon}_{\gamma}(u^{\mathcal{T}}), v\rangle + \left\langle \nabla\widehat{\psi}^{\mathcal{T}}_{\theta}(u^{\mathcal{T}}), v\right\rangle - b^{\epsilon}(\tilde{w}^{\mathcal{T}}, v) &= (f^{\epsilon}, v) && \forall v \in \mathcal{S}(\mathcal{T}),\\ -b^{\epsilon}(u^{\mathcal{T}}, v) - \Delta t\left(\nabla\tilde{w}^{\mathcal{T}}, \nabla v\right) &= -\left(u_{\text{old}}, v\right) && \forall v \in \mathcal{S}(\mathcal{T}).\end{aligned}$$

The new formulations allow to show uniqueness of the chemical potential $w^{\mathcal{T}}$.

Theorem 3.15. *Let γ and u_{old} satisfy the assumptions of Theorem 3.13 and $\theta > 0$. Then the solution of Problem 3.14 or, equivalently, Problem 3.12 is unique.*

Proof. Since uniqueness of $u^{\mathcal{T}}$ and $\nabla w^{\mathcal{T}}$ was already proved we only have to show uniqueness of $\int_{\Omega} w^{\mathcal{T}}(x)dx$. Let $w^{\mathcal{T}}_1$ and $w^{\mathcal{T}}_2$ be two solutions. Inserting them into the first variational equation in Problem 3.14 and subtracting one from the other yields

$$\left(w^{\mathcal{T}}_1 - w^{\mathcal{T}}_2, v\right) = 0 \qquad \forall v \in \mathcal{S}(\mathcal{T}).$$

Testing with $v = 1$ provides the assertion. □

Theorem 3.16. *Let γ, u_{old} and Δt satisfy the assumptions of Theorem 3.14 and $\theta > 0$. Then the solution of Problem 3.15 or, equivalently, Problem 3.13 is unique.*

Proof. The proof is analogue to the one of Theorem 3.15. □

For $\theta = 0$ where $\widehat{\psi}_{\theta}$ is the obstacle potential we can no longer write the variational inequalities as equations. However, we can slightly simplify them by imposing the constraint $u^{\mathcal{T}} \in \mathcal{K} \cap \mathcal{S}(\mathcal{T})$ manually, leading to classic variational inequalities of the first kind.

Problem 3.16. *Find $(u^{\mathcal{T}}, w^{\mathcal{T}}) \in \mathcal{K} \cap \mathcal{S}(\mathcal{T}) \times \mathcal{S}(\mathcal{T})$ such that*

$$\begin{aligned}\left\langle \tilde{\mathcal{F}}_{\gamma}(u^{\mathcal{T}}), v - u^{\mathcal{T}}\right\rangle - \left(w^{\mathcal{T}}, v - u^{\mathcal{T}}\right) &\geq \left(\tilde{f}, v - u^{\mathcal{T}}\right) && \forall v \in \mathcal{K} \cap \mathcal{S}(\mathcal{T}),\\ -\left(u^{\mathcal{T}}, v\right) - \Delta t\left(\nabla w^{\mathcal{T}}, \nabla v\right) &= -\left(u_{\text{old}}, v\right) && \forall v \in \mathcal{S}(\mathcal{T}).\end{aligned}$$

Problem 3.17. *Find $(u^{\mathcal{T}}, \tilde{w}^{\mathcal{T}}) \in \mathcal{K} \cap \mathcal{S}(\mathcal{T}) \times \mathcal{S}(\mathcal{T})$ such that*

$$\begin{aligned}\langle \mathcal{F}^{\epsilon}_{\gamma}(u^{\mathcal{T}}), v - u^{\mathcal{T}}\rangle - b^{\epsilon}(\tilde{w}^{\mathcal{T}}, v - u^{\mathcal{T}}) &\geq \left(f^{\epsilon}, v - u^{\mathcal{T}}\right) && \forall v \in \mathcal{K} \cap \mathcal{S}(\mathcal{T}),\\ -b^{\epsilon}(u^{\mathcal{T}}, v) - \Delta t\left(\nabla\tilde{w}^{\mathcal{T}}, \nabla v\right) &= -\left(u_{\text{old}}, v\right) && \forall v \in \mathcal{S}(\mathcal{T}).\end{aligned}$$

Theorem 3.17. *Let γ and u_{old} satisfy the assumptions of Theorem 3.13 and $\theta = 0$. Furthermore, assume that there is a $p \in \mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T})$ such that $|u^{\mathcal{T}}(p)| < 1$. Then the solution of Problem 3.16 or, equivalently, Problem 3.12 is unique.*

Proof. Again we only have to show uniqueness of $\int_{\Omega} w^{\mathcal{T}}(x)\,dx$. To this end let $w^{\mathcal{T}}_1$ and $w^{\mathcal{T}}_2$ be two solutions and let λ_p be the conforming nodal basis function associated with the vertex p. Then we have $u \pm \delta\lambda_p \in \mathcal{K} \cap \mathcal{S}(\mathcal{T})$ for all $\delta \in [0, \min\{1 - u^{\mathcal{T}}(p), u^{\mathcal{T}}(p) + 1\}]$. Testing the variational inequality with $u \pm \delta\lambda_p$ gives

$$\left\langle \tilde{\mathcal{F}}_{\gamma}(u^{\mathcal{T}}), \pm\delta\lambda_p\right\rangle - \left(w^{\mathcal{T}}_i, \pm\delta\lambda_p\right) \geq \left(\tilde{f}, \pm\delta\lambda_p\right).$$

Hence we even have

$$\left\langle \tilde{\mathcal{F}}_\gamma(u^{\mathcal{T}}), \lambda_p \right\rangle - \left(w_i^{\mathcal{T}}, \lambda_p\right) = \left(\tilde{f}, \lambda_p\right).$$

Subtracting this equation for $w_1^{\mathcal{T}}$ from the one for $w_2^{\mathcal{T}}$ and using the fact that $w_1^{\mathcal{T}} - w_2^{\mathcal{T}}$ is constant yields

$$0 = (w_1^{\mathcal{T}} - w_2^{\mathcal{T}}, \lambda_p) = \int_\Omega w_1^{\mathcal{T}}(x) - w_2^{\mathcal{T}}(x)\, dx \, \frac{(1, \lambda_p)}{|\Omega|}$$

and hence $w_1^{\mathcal{T}} - w_2^{\mathcal{T}} = 0$. □

Theorem 3.18. *Let γ, u_{old} and Δt satisfy the assumptions of Theorem 3.14 and $\theta > 0$. Furthermore, assume that there is a $p \in \mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T})$ such that $|u^{\mathcal{T}}(p)| < 1$. Then the solution of Problem 3.17 or, equivalently, Problem 3.13 is unique.*

Proof. The proof is analogue to the one of Theorem 3.17. □

3.4.4 Algebraic Formulation

In order to discuss the algebraic solution of the discrete Problems 3.12 and 3.13 we will now rewrite them in algebraic form as problems in $\mathbb{R}^n$. To this end we represent finite element functions $v \in \mathcal{S}(\mathcal{T})$ in terms of the conforming nodal basis $\mathcal{B}(\mathcal{T})$ which results in

$$v = \sum_{p \in \mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T})} v(p) \lambda_p.$$

Using an enumeration $\mathcal{B}(\mathcal{T}) = \{\lambda_{p_1}, \dots, \lambda_{p_n}\}$ with $n = |\mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T})|$ we can express finite element functions $v \in \mathcal{S}(\mathcal{T})$ by coefficient vectors $\underline{v} \in \mathbb{R}^n$ with $\underline{v}_i = v(p_i)$. The discrete analogues of the linear operators and right hand sides of Problem 3.12 in terms of coefficients are then given by

$$\begin{aligned}
&B \in \mathbb{R}^{n,n}, && B_{ij} = -\left(\lambda_{p_j}, \lambda_{p_i}\right), \\
&C \in \mathbb{R}^{n,n}, && C_{ij} = \Delta t \left(\nabla \lambda_{p_j}, \nabla \lambda_{p_i}\right), \\
&\underline{f} \in \mathbb{R}^{n}, && \underline{f}_i = \left(\tilde{f}, \lambda_{p_i}\right), \\
&\underline{g} \in \mathbb{R}^{n}, && \underline{g}_i = -\left(u_{\text{old}}, \lambda_{p_i}\right).
\end{aligned}$$

The nonlinear operator $\tilde{\mathcal{F}}_\gamma$ is represented by

$$F_\gamma : \mathbb{R}^n \to \mathbb{R}^n, \qquad F_\gamma(\underline{v})_i = \left\langle \tilde{\mathcal{F}}_\gamma(v), \lambda_{p_i} \right\rangle.$$

Finally the discrete analogue of the nonsmooth nonlinearity is given by

$$\varphi_\theta : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}, \qquad \varphi_\theta(\underline{v}) = \widehat{\psi}_\theta^{\mathcal{T}}(v) = \sum_{i=1}^n \widehat{\Psi}_\theta(\underline{v}_i) \int_\Omega \lambda_{p_i}(x) dx.$$

Now Problem 3.12 can be written as the operator inclusion:

Problem 3.18. *Find* $(\underline{u}^{\mathcal{T}}, \underline{w}^{\mathcal{T}}) \in \mathbb{R}^n \times \mathbb{R}^n$ *such that*

$$\begin{pmatrix} F_\gamma + \partial\varphi_\theta & B \\ B & -C \end{pmatrix} \begin{pmatrix} \underline{u}^{\mathcal{T}} \\ \underline{w}^{\mathcal{T}} \end{pmatrix} \ni \begin{pmatrix} \underline{f} \\ \underline{g} \end{pmatrix}.$$

If we additionally define

$$\begin{aligned} B^\epsilon &\in \mathbb{R}^{n,n}, & B^\epsilon_{ij} &= -\, b^\epsilon(\lambda_{p_j}, \lambda_{p_i}), \\ \underline{f}^\epsilon &\in \mathbb{R}^n, & \underline{f}^\epsilon_i &:= \frac{\theta_c + \epsilon}{2}(u_{\text{old}}, \lambda_{p_i}), \\ F^\epsilon_\gamma &: \mathbb{R}^n \to \mathbb{R}^n, & F^\epsilon_\gamma(\underline{v})_i &= \langle \tilde{\mathcal{F}}^\epsilon_\gamma(v), \lambda_{p_i} \rangle, \end{aligned}$$

then Problem 3.13 can be written as the operator inclusion:

Problem 3.19. *Find* $(\underline{u}^{\mathcal{T}}, \underline{\tilde{w}}^{\mathcal{T}}) \in \mathbb{R}^n \times \mathbb{R}^n$ *such that*

$$\begin{pmatrix} F^\epsilon_\gamma + \partial\varphi_\theta & B^\epsilon \\ B^\epsilon & -C \end{pmatrix} \begin{pmatrix} \underline{u}^{\mathcal{T}} \\ \underline{\tilde{w}}^{\mathcal{T}} \end{pmatrix} \ni \begin{pmatrix} \underline{f}^\epsilon \\ \underline{g} \end{pmatrix}.$$

Both problems are strongly nonlinear saddle point problems, and the algebraic solution will in general not be an easy task. Chapter 5 is dedicated to the development of fast and globally convergent algebraic solvers for nonsmooth nonlinear saddle point problems of this type. Before considering the whole problem we concentrate on nonsmooth nonlinear minimization problems without equality constraints in the next chapter. Their efficient solution will be needed in the algorithm for saddle point problems developed later on.

4 Truncated Nonsmooth Newton Multigrid for Nonsmooth Convex Minimization Problems

In this chapter we consider the fast algebraic solution of finite-dimensional nonlinear convex minimization problems with a nonlinearity consisting of a smooth part and a nonsmooth part which is given componentwise. While this problem class includes component wise inequality constraints, i.e. box-constraint, no linear constraints are present here. Since the presented algorithms will depend in general on the selection of a basis in a finite-dimensional space we will introduce them in $\mathbb{R}^n$.

We start by introducing the class of minimization problems. Then we analyze the classical nonlinear Gauß–Seidel and Jacobi methods for this problem class and introduce associated nonsmooth Newton methods. Finally we propose a multigrid method for the linear subproblems. Application of a fixed number of multigrid steps to each subproblem then leads to an overall nonlinear multigrid method that converges globally for the considered problem class.

4.1 Nonsmooth Convex Minimization Problems

We consider the nonlinear minimization problem

$$u^* \in V : \qquad J(u^*) \leq J(v) \qquad \forall v \in V \tag{4.1}$$

where V is a finite-dimensional vector space and

$$J = J_0 + \varphi : V \to \mathbb{R} \cup \{\infty\} \tag{4.2}$$

is convex and lower semicontinuous, i.e.,

$$v^k \to v \Rightarrow J(v) \leq \liminf_{k \to \infty} J(v^k).$$

Assuming that a basis of V is fixed we will consider $V = \mathbb{R}^n$ in the following. Regarding the smooth part J_0 we need the following assumptions:

(A3) $J_0 : \mathbb{R}^n \to \mathbb{R}$ is strictly convex and continuously differentiable. Its derivative ∇J_0 is Lipschitz continuous with the Lipschitz constant $L_{\nabla J_0}$ and there is a symmetric positive definite matrix $\underline{H}_{J_0} \in \mathbb{R}^{n,n}$ such that for all $u, v \in \mathbb{R}^n$ we have

$$\langle \nabla J_0(u) - \nabla J_0(v), u - v \rangle \geq \left\langle \underline{H}_{J_0}(u - v), u - v \right\rangle . \tag{4.3}$$

On the one hand we do not want to assume smoothness of ∇J_0. On the other hand we want to develop Newton-type methods for this problem class. Hence we need some concept of a generalized Hessian of J_0.

By Rademacher's theorem (see, e.g., [81]) Lipschitz continuous operators $T : \mathbb{R}^n \to \mathbb{R}^m$ are differentiable almost everywhere and the set

$$\mathcal{D}_T := \{u \in \mathbb{R}^n : T \text{ is differentiable in } u\} \tag{4.4}$$

is dense in $\mathbb{R}^n$. Thus we can define the following generalized derivatives.

Definition 4.1. *For a Lipschitz continuous operator $T : \mathbb{R}^n \to \mathbb{R}^m$ the B-subdifferential $\partial_B T$ (cf. [88, 106]) and the generalized Jacobian in the sense of Clarke $\partial_C T$ (cf. [34]) are defined by*

$$\partial_C T(u) := \operatorname{co} \partial_B T(u), \qquad \partial_B T(u) := \Big\{ \lim_{k\to\infty} \nabla T(u_k) : u_k \to u, u_k \in \mathcal{D}_T \Big\}.$$

While the generalized Jacobian is often denoted by ∂ we denote it by ∂_C to distinguish it from other generalized linearizations that will be denoted by ∂. For convex functionals f the symbol ∂f denotes the usual subdifferential. Furthermore, $\partial^2 f$ will denote a generalized linearization of the gradient ∇f or the subdifferential ∂f and thus a generalized Hessian of f.

(A4) $J_0 : \mathbb{R}^n \to \mathbb{R}$ satisfies (A3). For each $u \in \mathbb{R}^n$ there is a symmetric positive definite matrix $\partial^2 J_0(u) \in \mathbb{R}^{n,n}$ representing a generalized linearization of $\nabla J_0(u)$ that satisfies

$$\langle \underline{H}_{J_0} v, v \rangle \leq \langle \partial^2 J_0(u) v, v \rangle \qquad \forall u, v \in \mathbb{R}^n. \tag{4.5}$$

If ∇J_0 is differentiable everywhere $\partial^2 J_0 = \nabla^2 J_0$ is chosen.

Later on we will also need to assume boundedness of the generalized Hessian:

(A5) $J_0 : \mathbb{R}^n \to \mathbb{R}$ satisfies (A3) and (A4). Furthermore, there is a symmetric positive definite matrix $\overline{H}_{J_0} \in \mathbb{R}^{n,n}$ such that

$$\langle \partial^2 J_0(u) v, v \rangle \leq \langle \overline{H}_{J_0} v, v \rangle \qquad \forall u, v \in \mathbb{R}^n. \tag{4.6}$$

If ∇J_0 is differentiable we simply have $\partial^2 J_0(u) = \nabla^2 J_0(u)$ in (A4). In the general case the application of Lemma A.3 to $T = \nabla J_0$ still guarantees the strong monotonicity (4.3) of ∇J_0 in (A3) if $\partial^2 J_0(u) \in \partial_C(\nabla J_0)(u)$ and if the Hessian of J_0 is bounded from below, where it exists, i.e., if there is a symmetric positive definite matrix $\underline{H}_{J_0} \in \mathbb{R}^{n,n}$ such that

$$\langle \underline{H}_{J_0} v, v \rangle \leq \langle \nabla^2 J_0(u) v, v \rangle \qquad \forall u \in \mathcal{D}_{\nabla J_0}, v \in \mathbb{R}^n.$$

The most important example for smooth convex functions J_0 are strictly convex quadratic functions, i.e., functions satisfying

(A6) $J_0 : \mathbb{R}^n \to \mathbb{R}$ is given by

$$J_0(u) = \frac{1}{2}\langle Au, u\rangle - \langle b, u\rangle$$

with a symmetric positive definite matrix $A \in \mathbb{R}^{n,n}$ and $b \in \mathbb{R}^n$.

Assumption (A6) directly implies (A5) with $\underline{H}_{J_0} = \overline{H}_{J_0} = \partial^2 J_0(u) = \nabla^2 J_0(u) = A$. Another special case are functions satisfying

(A7) $J_0 : \mathbb{R}^n \to \mathbb{R}$ takes the form

$$J_0(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle + \sum_{i=1}^{\overline{n}} \gamma_i^2(D_i v)$$

with a symmetric positive semidefinite matrix $A \in \mathbb{R}^{n,n}$, matrices $D_i \in \mathbb{R}^{d,n}$, and continuously differentiable convex functions $\gamma_i^2 : \mathbb{R}^d \to \mathbb{R}$ with Lipschitz continuous derivatives. For each $i = 1, \dots, \overline{n}$ there are symmetric positive semidefinite matrices $\underline{H}_{\gamma_i^2}, \overline{H}_{\gamma_i^2} \in \mathbb{R}^{d,d}$ such that for $\partial^2(\gamma_i^2)(x) \in \partial_C(\nabla\gamma_i^2)(x)$ we have

$$\left\langle \underline{H}_{\gamma_i^2} v, v\right\rangle \leq \left\langle \partial^2(\gamma_i^2)(y)v, v\right\rangle \leq \left\langle \overline{H}_{\gamma_i^2} v, v\right\rangle \qquad \forall u, v \in \mathbb{R}^d. \tag{4.7}$$

To simplify notation we denote by $D \in (\mathbb{R}^d)^{\overline{n},n} = (\mathbb{R}^{d,1})^{\overline{n},n}$ the block matrix whose entries are the column vectors $(D_i)_j \in \mathbb{R}^d$. Then its transposed matrix can be regarded to be $D^T \in (\mathbb{R}^{1,d})^{n,\overline{n}}$. Furthermore, we use the block diagonal matrices $\partial^2(\gamma^2)(x), \underline{H}_{\gamma^2}, \overline{H}_{\gamma^2} \in (\mathbb{R}^{d,d})^{\overline{n},\overline{n}}$ defined by

$$\partial^2(\gamma^2)(x) = \operatorname{diag} \partial^2(\gamma_i^2)(x_i), \quad \underline{H}_{\gamma^2} = \operatorname{diag} \underline{H}_{\gamma_i^2}, \quad \overline{H}_{\gamma^2} = \operatorname{diag} \overline{H}_{\gamma_i^2}.$$

Using this notation $A + D^T \underline{H}_{\gamma^2} D$ is assumed to be symmetric positive definite.

Functions of this type are especially interesting since the smooth convex part of the energy functionals associated with the discrete saddle point problems in Section 3.4.4 take this form. There, each D_i is the matrix computing the gradient of a finite element function on the i-th grid element τ_i, while γ_i^2 is the anisotropy γ^2 scaled by the element integration weight $\int_{\tau_i} 1$. The requirements on the γ_i^2 are then ensured if γ satisfies (A1) and (A2).

Lemma 4.1. *(A7) implies (A5) with*

$$\partial^2 J_0(u) = A + D^T \partial^2(\gamma^2)(Du)D, \quad \underline{H}_{J_0} = A + D^T \underline{H}_{\gamma^2} D, \quad \overline{H}_{J_0} = A + D^T \overline{H}_{\gamma^2} D.$$

Proof. (4.5) and (4.6) follow directly from (4.7) and the representation

$$A + D^T M D = A + \sum_{i=1}^{\overline{n}} D_i^T M_{ii} D_i$$

for $M = \partial(\gamma^2)(v), \underline{H}_{\gamma^2}, \overline{H}_{\gamma^2}$. Elementary computations provide

$$\nabla J_0(v) = Av - b + (\nabla\gamma^2)(Dv)D = Av - b + \sum_{i=1}^{\overline{n}} \nabla\gamma_i^2(D_i v)D_i.$$

Application of Lemma A.3 to $\nabla\gamma_i^2$ together with (4.7) now implies (4.3). □

While J_0 is assumed to satisfy certain smoothness properties, φ represents the nonsmooth part of J at the price that this nonlinearity decouples with respect to the unknowns.

(A8) $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ takes the form

$$\varphi(v) = \sum_{i=1}^{n} \varphi_i(v_i).$$

Each $\varphi_i : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is convex, lower semicontinuous on $\mathbb{R}$, continuous on its domain $\operatorname{dom}\varphi_i$, and twice continuously differentiable on a finite number of disjoint nonempty open intervals (a_i^k, a_i^{k+1}), $a_i^k \in \mathbb{R} \cup \{-\infty, +\infty\}$ having the property

$$\overline{\operatorname{dom}\varphi_i} = \overline{\{x : \varphi_i(x) < \infty\}} = \bigcup\nolimits_{k=1}^{m_i} \overline{(a_i^{k-1}, a_i^k)} = \overline{(a_i^0, a_i^{m_i})}.$$

The intervals are maximal in the sense that φ_i is not twice continuously differentiable on (a_i^k, a_i^{k+2}). Furthermore, the limits

$$\lim_{\xi \nearrow a_i^{k+1}} \varphi_i''(\xi), \qquad \lim_{\xi \searrow a_i^k} \varphi_i''(\xi)$$

exist in $\mathbb{R} \cup \{\infty\}$ for $k = 0, \ldots, (m_i - 1)$.

Under these assumptions existence, uniqueness and stability of solutions follow using standard arguments:

Proposition 4.1. *Assume that (A3) and (A8) hold. Then J is strictly convex, proper, lower semicontinuous and coercive. The subdifferential $\partial J : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ has a single-valued Lipschitz continuous inverse $(\partial J)^{-1} : \mathbb{R}^n \to \mathbb{R}^n$.*

Proof. Lower semicontinuity follows directly from the regularity assumptions on J_0 and φ_i. The fundamental theorem of calculus and (4.3) imply the strong convexity

$$J_0(u) - J_0(v) \geq \frac{1}{2}\left\langle \underline{H}_{J_0}(u - v), u - v \right\rangle + \left\langle \nabla J_0(v), u - v \right\rangle. \tag{4.8}$$

and thus strict convexity of J_0.

By definition φ is finite and continuously differentiable in at least one point $\tilde{u}$. Hence φ is bounded from below by the affine function $\partial\varphi(\tilde{u})$. This and the above inequality with $v = 0$ also implies $J(u) \to \infty$ for $\|u\| \to \infty$. Thus $J(\cdot) - \langle y, \cdot \rangle$ has a unique minimizer $x = (\partial J)^{-1}(y)$ (see [49, Chapter II, Proposition 1.2]) which is the unique solution of the variational inequality (see [49, Chapter II, Proposition 2.2])

$$x \in \mathbb{R}^n : \qquad \langle \nabla J_0(x), v - x \rangle + \varphi(v) - \varphi(x) \geq \langle y, v - x \rangle \qquad \forall v \in \mathbb{R}^n. \tag{4.9}$$

This provides single-valuedness of the inverse operator $(\partial J)^{-1}$. Now let $x_i = (\partial J)^{-1}(y_i)$, $i = 1, 2$. Then testing the inequality for i with $j \neq i$ leads to the Lipschitz continuity

$$\begin{aligned} \|x_1 - x_2\|^2_{\underline{H}_{J_0}} &\leq \langle \nabla J_0(x_1) - \nabla J_0(x_2), x_1 - x_2 \rangle \\ &\leq \langle y_1 - y_2, x_1 - x_2 \rangle \leq \|x_1 - x_2\|_{\underline{H}_{J_0}} \|y_1 - y_2\|_{\underline{H}^{-1}_{J_0}}. \end{aligned}$$

□

If a minimization problem is constrained to a convex set K this can in general be incorporated by adding the indicator functional

$$\chi_K : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}, \qquad \chi_K(x) = \begin{cases} 0 & \text{if } x \in K \\ \infty & \text{else} \end{cases}$$

to the function to minimize. The convex sets leading to indicator functions χ_K satisfying (A8) are possibly unbounded hypercubes as given in the following example.

Example 4.1. *Let $\underline{\psi} \in (\mathbb{R} \cup \{-\infty\})^n$, $\overline{\psi} \in (\mathbb{R} \cup \{\infty\})^n$ with $\underline{\psi} \leq \overline{\psi}$. Then*

$$K := \{u \in \mathbb{R}^n : \underline{\psi} \leq u \leq \overline{\psi}\}$$

is closed and convex and $\varphi = \chi_K$ satisfies (A8). If additionally J_0 satisfies (A6) the minimization problem for $J = J_0 + \varphi$ is a discrete quadratic obstacle problem.

4.2 Truncated Nonsmooth Newton Methods for Convex Minimization Problems

For the rest of this chapter we consider the algebraic solution of the minimization problem (4.1) for J given by (4.2) assuming that J_0 and φ satisfy (A3) and (A8), respectively. Since we are especially interested in problems obtained by discretization of partial differential equations, we will now introduce nonlinear monotone iterative methods that can be combined with multigrid techniques to obtain fast and globally convergent nonlinear multigrid methods.

While many common methods for nonsmooth minimization directly apply multilevel ideas to the minimization problem [69, 80, 104] we will present a different approach that is based on a Newton-type method for a related Lipschitz continuous operator. This approach can also be viewed as a globally convergent extension of the well-known primal and primal–dual active set methods for obstacle problems [14, 64, 65].

4.2.1 Monotone Nonlinear Smoothers

Simple linear iterations for linear system with a symmetric positive definite matrix can often be expressed in terms of minimization algorithms for the associated minimization problem. For example the Gauß–Seidel iteration is equivalent to successive minimization in the coordinate directions while the Jacobi iteration is equivalent to parallel minimization in these directions.

Based on this interpretation the extension of these methods to nonsmooth nonlinear minimization problems is straightforward. We will describe these iterations in terms of correction operators $\mathcal{F}$ with

$$u^{\nu+1} = u^{\nu} + \mathcal{F}(u^{\nu}). \tag{4.10}$$

The nonlinear Gauß–Seidel method $\mathcal{F}_{GS}$ obtained by successive minimization of J in the coordinate directions e_i can be written as

$$\mathcal{F}_{GS}(x)_i = \rho\biggl(J, x + \sum_{j=1}^{i-1} \mathcal{F}_{GS}(x)_j e_j, e_i\biggr).$$

using the notation

$$\rho(J, x, d) := \operatorname*{arg\,min}_{\rho \in \mathrm{dom}(J)} J(x + \rho d) \tag{4.11}$$

for the minimizer of J along a line $x + \mathbb{R}d$. In the case that J_0 satisfies (A6), i.e., that it is a quadratic energy, this reduces to

$$\mathcal{F}_{GS}(x) = (D + L + \partial\varphi)^{-1}(b - Rx) - x \tag{4.12}$$

for the splitting $A = D + L + R$ of A in diagonal, left and right part.

Similarly the nonlinear Jacobi method $\mathcal{F}_J$ obtained by parallel minimization of J in the coordinate directions e_i can be written as

$$\mathcal{F}_J(x)_i = \rho(J, x, e_i).$$

which reduces for the quadratic energy to

$$\mathcal{F}_J(x) = (D + \partial\varphi)^{-1}(b - (L + R)x) - x. \tag{4.13}$$

Proposition 4.2. *$\mathcal{F}_{GS}$ and $\mathcal{F}_J$ are well-defined, Lipschitz continuous operators having the property*

$$\mathcal{F}_{GS}(x) = \mathcal{F}_J(x) = 0 \qquad \Leftrightarrow \qquad \partial J(x) = \nabla J_0(x) + \partial\varphi(x) \ni 0,$$

i.e., the correction at x is zero if and only if x solves (4.1).

Proof. First we note that $\mathcal{F}_{GS}$ can be evaluated successively starting from the first component by solving scalar minimization problems for convex functionals $J_0(y+(\cdot)e_i)+\varphi_i(\cdot)$ for varying $y \in \mathbb{R}^n$. For $\mathcal{F}_J$ this is even possible in parallel. By standard arguments these minimization problems are equivalent to variational inequalities

$$r \in \mathbb{R}: \qquad \langle \nabla J_0(y+re_i),(v-r)e_i\rangle + \varphi_i(y_i+v) - \varphi_i(y_i+r) \geq 0 \qquad \forall v \in \mathbb{R} \quad (4.14)$$

which have a unique solution by the same arguments as used in Proposition 4.1.

In contrast to Proposition 4.1 the datum y appears in the nonlinear operator here. Thus we have to extend the arguments in order to prove Lipschitz continuity. To this end let $r^1, r^2 \in \mathbb{R}$ be two solutions of (4.14) for data $y^1, y^2 \in \mathbb{R}^n$, respectively. Without loss of generality assume that $y_i^1 = y_i^2 = 0$. Testing the variational inequality for r^1 with r^2 and vice versa yields

$$\left\langle \nabla J_0(v^2) - \nabla J_0(v^1), (r^2-r^1)e_i \right\rangle \leq 0$$

for $v^j = y^j + r^j e_i$. Together with (A3) this gives

$$\begin{aligned} \|v^2-v^1\|^2_{\underline{H}_{J_0}} &\leq \left\langle \nabla J_0(v^2) - \nabla J_0(v^1), v^2-v^1 \right\rangle \\ &\leq \left\langle \nabla J_0(v^2) - \nabla J_0(v^1), y^2-y^1 \right\rangle \\ &\leq L_{\nabla J_0}\|v^2-v^1\|\|y^2-y^1\| \end{aligned}$$

and thus

$$|r^2-r^1| \leq \|v^2-v^1\| \leq \frac{\|v^2-v^1\|^2_{\underline{H}_{J_0}}}{\lambda_{\min}(\underline{H}_{J_0})\|v^2-v^1\|} \leq \frac{L_{\nabla J_0}}{\lambda_{\min}(\underline{H}_{J_0})}\|y^2-y^1\|.$$

Hence $\mathcal{F}_J(x)_i$ and $\mathcal{F}_{GS}(x)_i$ depend Lipschitz continuously on x and the latter also on $\mathcal{F}_{GS}(x)_j$, $j<i$, showing the Lipschitz continuity of both operators. Now assume that $\mathcal{F}_{GS}(x)=0$ or $\mathcal{F}_J(x)=0$. Then

$$\langle \nabla J_0(x),(v-x)_i e_i\rangle + \varphi_i(v_i) - \varphi_i(x_i) \geq 0 \qquad \forall v_i \in \mathbb{R}$$

holds for $i=1,\dots,n$. Summing up these inequalities yields (4.9) with $y=0$, which is equivalent to (4.1). □

This result is a key ingredient to prove the convergence of these iterative methods. The following convergence result was proved by Kornhuber [69, 72] for quadratic obstacle problems and piecewise quadratic problems. In order to extend the proof to the more general case considered here we introduce the following monotonicity assumption for a correction operator $\mathcal{F}$:

(A9) The operator $\mathcal{F} : \operatorname{dom}\varphi \to \mathbb{R}^n$ is monotone in the sense that $J(u+\mathcal{F}(u)) \leq J(u)$ holds for all $u \in \operatorname{dom}\varphi$. If a sequence $u^\nu \in \operatorname{dom}\varphi$ satisfies $u^\nu \to u'$ and

$$J(u^{\nu+1}) \leq J(u^\nu + \mathcal{F}(u^\nu)) \qquad \forall \nu \in \mathbb{N} \qquad (4.15)$$

then $u' = u^*$ is the minimizer of J.

Under this assumption we can even relax the requirements on J:

Theorem 4.1. *Assume that J is strictly convex, lower semicontinuous, coercive and proper and that $\mathcal{F}$ satisfies (A9). Furthermore, assume that $J(x+\mathcal{C}(x)) \leq J(x)$. Then the sequence u^ν generated by*

$$\begin{aligned} u^{\nu+\frac{1}{2}} &= u^\nu + \mathcal{F}(u^\nu), \\ u^{\nu+1} &= u^{\nu+\frac{1}{2}} + \mathcal{C}(u^{\nu+\frac{1}{2}}) \end{aligned} \tag{4.16}$$

converges to the unique solution u^* of (4.1) *for every $u^0 \in \operatorname{dom}\varphi$.*

Proof. By monotonicity of $\mathcal{F}$ and $\mathcal{C}$ and the coercivity of J the sequence (u^ν) is bounded. Thus there exists a convergent subsequence (u^{ν_i}). Now let (u^{ν_i}) be any convergent subsequence with $u^{\nu_i} \to u'$. Then we have

$$J(u^{\nu_{i+1}}) \leq J(u^{\nu_i+1}) \leq J(u^{\nu_i} + \mathcal{F}(u^{\nu_i})) \leq J(u^{\nu_i})$$

and from (A9) we get $u' = u^*$. Thus each accumulation point u' must be the solution of (4.9) which provides convergence of the whole sequence. □

Corollary 4.1. *The Gauß–Seidel method $\mathcal{F}_{GS}$ satisfies assumption (A9) and the sequence generated by* (4.16) *with $\mathcal{F} = \mathcal{F}_{GS}$ converges to the unique solution u^* of* (4.1) *for every $u^0 \in \operatorname{dom}\varphi$.*

Proof. Monotonicity of $\mathcal{F}_{GS}$ is given by construction. Now let $u^\nu \in \operatorname{dom}\varphi$ be a convergent sequence with $u^\nu \to u'$ and (4.15). Then we have

$$J(u^{\nu+1}) \leq J(u^\nu + \mathcal{F}_{GS}(u^\nu)) \leq J(u^\nu) \qquad \forall \nu \in \mathbb{N}$$

and thus by continuity of $\mathcal{F}_{GS}$ and J on $\operatorname{dom}\varphi$

$$J(u') = \lim_{\nu\to\infty} J(u^\nu + \mathcal{F}(u^\nu)) = J(u' + \mathcal{F}(u')).$$

Since the Gauß–Seidel method implies $\mathcal{F}_{GS}(u) = 0$ if $J(u + \mathcal{F}_{GS}(u)) = J(u)$ holds, this yields $\mathcal{F}_{GS}(u') = 0$ and, by Proposition 4.2, $u' = u^*$. Hence Theorem 4.1 can be applied to obtain global convergence. □

The essential property used to prove convergence of the Gauß–Seidel method is

$$\mathcal{F}_{GS}(x) \neq 0 \qquad \Rightarrow \qquad J(x + \mathcal{F}_{GS}(x)) < J(x). \tag{4.17}$$

While the Jacobi method does in general not have this property, the component wise corrections $x + (\mathcal{F}_J(x) - x)_i e_i$ still satisfy the following monotonicity by construction:

$$\mathcal{F}_J(x) \neq 0 \qquad \Rightarrow \begin{cases} \forall i: & J(x + \mathcal{F}_J(x)_i e_i) \leq J(x), \\ \exists i': & J(x + \mathcal{F}_J(x)_{i'} e_{i'}) < J(x). \end{cases} \tag{4.18}$$

As it is known from the Jacobi method for linear problems we will need to introduce damping to ensure convergence. To this end we define the directional sublevel set of J along the line $x + \mathbb{R}d$ by

$$\rho_<(J, x, d, \overline{r}) := \{r : J(x + rd) \leq J(x + \overline{r}d)\}.$$

Then we clearly have $\rho(J, x, d) \in \rho_<(J, x, d, \overline{r})$ for all $\overline{r} \in \mathbb{R}$.

Corollary 4.2. *The damped Jacobi iteration* $\mathcal{F}(x) = \rho_J(x)\mathcal{F}_J(x)$ *with*

$$\rho_J(x) \in \rho_< \left(J, x, \mathcal{F}_J(x), \frac{1}{n}\right) \tag{4.19}$$

satisfies assumption (A9) and the sequence generated by (4.16) *with* $\mathcal{F} = \rho_J\mathcal{F}_J$ *converges to the unique solution* u^* *of* (4.1) *for every* $u^0 \in \operatorname{dom}\varphi$.

Proof. Since $x + \frac{1}{n}\mathcal{F}_J(x)$ is a convex combination of the component wise corrections $x + \mathcal{F}_J(x)_i e_i$ the monotonicity and property (4.17) for the Lipschitz continuous operator $\frac{1}{n}\mathcal{F}_J$ follows from (4.18) and convexity of J. Monotonicity of $\rho_J\mathcal{F}_J$ is then given by (4.19). Now let $u^\nu \in \operatorname{dom}\varphi$ be a convergent sequence with $u^\nu \to u'$ and (4.15). Then we have

$$\begin{aligned} J(u^{\nu+1}) &\leq J(u^\nu + \rho_J(u^\nu)\mathcal{F}_J(u^\nu)) \\ &\leq J\left(u^\nu + \frac{1}{n}\mathcal{F}_J(u^\nu)\right) \leq J(u^\nu) \qquad \forall \nu \in \mathbb{N}. \end{aligned}$$

Now we can proceed literally as in Corollary 4.1 with $\mathcal{F}_{GS}$ replaced by $\frac{1}{n}\mathcal{F}_J$. □

The cheaply available damping parameter $\rho_J(u^\nu) = 1/n$ will in general be much too pessimistic. More reasonable damping parameters can be obtained by approximating the exact minimizer $\rho(J, u^\nu, \mathcal{F}_J(u^\nu))$ starting from $1/n$, e.g. by bisection.

4.2.2 Inexact Nonlinear Smoothers

Until now we implicitly assumed that the solution of the scalar minimization problems within one step of the Gauß–Seidel or Jacobi method can be done exactly. While this is true for quadratic or piecewise quadratic J it does in general not hold for other nonlinear energies. In this subsection we give a convergence proof that also holds true for inexact evaluation of these minimization problems e.g. by bisection. To prove the convergence result we will extend the ideas of the proof given in Kornhuber [73] for the Gauß–Seidel method and a quadratic J_0 to more general energies.

Theorem 4.2. *Let* $\overline{\mathcal{F}}_{GS}$ *be an inexact version of* $\mathcal{F}_{GS}$ *given by*

$$\overline{\mathcal{F}}_{GS}(x)_i = \omega(x,i)\rho\left(J, x + \sum_{j=1}^{i-1} \overline{\mathcal{F}}_{GS}(x)_j e_j, e_i\right), \qquad \omega(x,i) \in [\omega_0, 1]$$

for some fixed $\omega_0 \in (0, 1]$. *Then* $\overline{\mathcal{F}}_{GS}$ *satisfies (A9).*

Proof. Monotonicity of $\overline{\mathcal{F}}_{GS}$ is given by construction. Furthermore, by convexity, the inequality (4.8), and the inequality (4.14) with $v = 0$ we can estimate the norm of the correction $T_i y := \rho(J, y, e_i)e_i$ for any $y \in \operatorname{dom} \varphi$ by

$$\begin{aligned} J(y) - J(y + \omega_0 T_i y) &\geq \omega_0 \Big(J_0(y) - J_0(y + T_i y) + \varphi_i(y_i) - \varphi_i(y_i + T_i y) \Big) \\ &\geq \omega_0 \left(\frac{1}{2} \|T_i y\|^2_{\underline{H}_{J_0}} + \langle \nabla J_0(y + T_i y), -T_i y \rangle + \varphi_i(y_i) - \varphi_i(y_i + T_i y) \right) \\ &\geq \frac{\omega_0}{2} \|T_i y\|^2_{\underline{H}_{J_0}}. \end{aligned}$$

Now let $u^\nu \in \operatorname{dom} \varphi$ be a convergent sequence satisfying $u^\nu \to u'$ and (4.15). Set

$$w_i^\nu := u^\nu + \sum_{j=1}^{i} \overline{\mathcal{F}}_{GS}(x)_j e_j = u^\nu + \sum_{j=1}^{i} \omega(u^\nu, j) T_j w_{j-1}^\nu.$$

Then we have for all $\nu \in \mathbb{N}$

$$J(u^{\nu+1}) \leq J(w_i^\nu) \leq J(w_{i-1}^\nu + \omega_0 T_i w_{i-1}^\nu) \leq J(w_{i-1}^\nu) \leq J(u^\nu).$$

This monotonicity and the above estimate yield

$$J(u^\nu) - J(u^{\nu+1}) \geq J(w_{i-1}^\nu) - J(w_{i-1}^\nu + \omega_0 T_i w_{i-1}^\nu) \geq \frac{\omega_0}{2} \|T_i w_{i-1}^\nu\|^2_{\underline{H}_{J_0}}.$$

Now continuity of J gives $T_i w_{i-1}^\nu \to 0$. Together with the triangle inequality and $\omega(u^\nu, j) \leq 1$ this implies

$$\|w_i^\nu - u^\nu\|_{\underline{H}_{J_0}} = \left\| \sum_{j=1}^{i} \omega(u^\nu, j) T_j w_{j-1}^\nu \right\|_{\underline{H}_{J_0}} \leq \sum_{j=1}^{i} \|T_j w_{j-1}^\nu\|_{\underline{H}_{J_0}} \to 0$$

and thus $T_i u' = \lim_{\nu \to \infty} T_i w_{i-1}^\nu = 0$. Hence u' is a fixed point of $\mathcal{F}_{GS}$ and thus $u' = u^*$. □

Theorem 4.3. *Let $\overline{\mathcal{F}}_J$ be an inexact version of $\mathcal{F}_J$ given by*

$$\overline{\mathcal{F}}_J(x)_i = \omega(x, i) \rho\Big(J, x, e_i \Big), \qquad \omega(x, i) \in [\omega_0, 1]$$

for some fixed $\omega_0 \in (0, 1]$. Then the damped method $\rho_J(x) \overline{\mathcal{F}}_J(x)$ with

$$\rho_J(x) \in \rho_< \left(J, x, \overline{\mathcal{F}}_J(x), \frac{1}{n} \right)$$

satisfies assumption (A9).

Proof. By definition of ρ_J, convexity, and monotonicity we have with $T_i y := \rho(J,y,e_i)e_i$

$$J\Big(x + \rho_J(x)\overline{\mathcal{F}}_J(x)\Big) \leq J\Big(x + \frac{1}{n}\sum_{i=1}^{n} \omega(x,i)T_i x\Big) \leq \frac{1}{n}\sum_{i=1}^{n} J(x + \omega(x,i)T_i x)$$
$$\leq \frac{1}{n}\sum_{i=1}^{n} J(x + \omega_0 T_i x) \leq J(x).$$

This inequality and the first estimate in the proof of Theorem 4.2 provide

$$J(x) - J(x + \rho_J(x)\overline{\mathcal{F}}_J(x)) \geq \frac{1}{n}\sum_{i=1}^{n} J(x) - J(x + \omega_0 T_i x) \geq \frac{\omega_0}{2n}\sum_{i=1}^{n} \|T_i x\|^2_{\underline{H}_{J_0}}.$$

Now let $u^\nu \in \operatorname{dom}\varphi$ be a convergent sequence satisfying $u^\nu \to u'$ and (4.15). Then the last estimate implies

$$J(u^\nu) - J(u^{\nu+1}) \geq \frac{\omega_0}{2n}\sum_{i=1}^{n} \|T_i u^\nu\|^2_{\underline{H}_{J_0}} \to 0$$

and thus $T_i u' = \lim_{\nu\to\infty} T_i u^\nu = 0$. Hence u' is a fixed point of $\mathcal{F}_J$ and thus $u' = u^*$. □

Theorem 4.2 and Theorem 4.3 allow to replace $\mathcal{F}_{GS}$ and $\rho_J\mathcal{F}_J$ by their inexact versions $\overline{\mathcal{F}}_{GS}$ and $\rho_J\overline{\mathcal{F}}_J$ in the iteration in Theorem 4.1. A practical stopping criterion for the bisection method that guarantees the accuracy condition required in Theorem 4.2 and Theorem 4.3 is given in [75].

4.2.3 Derivatives of Nonlinear Smoothers

Although the Gauß–Seidel and the damped Jacobi method converge globally for the above minimization problem (4.1) they are in general not satisfactory. For the linear case where $J = J_0$ is quadratic it is well known that their convergence rates deteriorate rapidly if the problem results from a discretized differential operator. This property is in general directly inherited by the nonlinear versions.

However, the convergence results offer the possibility to introduce additional corrections $\mathcal{C}$ to accelerate the convergence. This is used in [69, 72, 75] to add corrections obtained by minimization in the direction of certain subspaces spanned by coarse grid functions. While this approach leads to a sequence of nonlinear subproblems we will present a simpler approach (first introduced in [58] for quadratic obstacle problems) that leads to a linear subproblem in each iteration step.

The usual application of Newton's method for a minimization problem would deal with the equivalent optimality system. Due to the nonsmooth nonlinearity the minimization problem (4.1) is in general only equivalent to the inclusion

$$\partial J(u^*) = \nabla J_0(u^*) + \partial\varphi(u^*) \ni 0 \tag{4.20}$$

where $\partial\varphi$ is the set-valued subdifferential of φ. The application of a classical Newton method would require ∂J to be differentiable. Since this operator is in general not even single-valued for the present problem class we have to look for a different formulation of (4.20).

The iteration (4.10) can be regarded as fixed point iteration for the operator $I + \mathcal{F}$ where I is the identity matrix. Instead of looking at the induced iteration we now consider the operator $\mathcal{F}$ for a reformulation of the problem. As long as $\mathcal{F}$ satisfies assumption (A9) the minimization problem (4.1) and the variational inclusion (4.20) are equivalent to

$$\mathcal{F}(u^*) = 0 \tag{4.21}$$

for the continuous operator $\mathcal{F}$. This offers the possibility to apply a Newton-type method to (4.21). Since $\mathcal{F}_{GS}$ and $\mathcal{F}_J$ are even Lipschitz continuous operators a standard nonsmooth Newton method as introduced in Kummer [79], Pang [85], Qi and Sun [89] using a generalized Jacobian in the sense of Clarke (see [34]) as linearization for $\mathcal{F}$ could be applied. However, it will not be an easy task in general to compute generalized derivatives since a classical chain rule does not hold for the considered operators. Thus we will postulate the chain rule and use it to construct a generalized linearization of $\mathcal{F}$ that will in general not coincide with the generalized Jacobian.

We restrict our considerations regarding linearizations to the case that J_0 satisfies (A6); i.e., the smooth part of J is quadratic. In this case $\mathcal{F}_{GS}$ and $\mathcal{F}_J$ take the form (4.12) and (4.13), respectively. The resulting algorithms will be extended to other smooth nonlinearities J_0 satisfying (A3) later on.

The derivation of linearizations of $f_i := (A_{ii}(\cdot) + \partial\varphi_i(\cdot))^{-1} : \mathbb{R} \to \mathbb{R}$ is essential for the construction of generalized linearizations of $\mathcal{F}_{GS}$ and $\mathcal{F}_J$. Thus we investigate the smoothness of the scalar functionals φ_i and their piecewise derivatives first.

Lemma 4.2. *The limits*

$$\begin{aligned}
\varphi'_{i,-}(x) &:= \lim_{\xi \nearrow x} \varphi'_i(\xi), &\qquad \varphi''_{i,-}(x) &:= \lim_{\xi \nearrow x} \varphi''_i(\xi) &\qquad \forall x \in (a_i^0, a_i^{m_i}],\\
\varphi'_{i,+}(x) &:= \lim_{\xi \searrow x} \varphi'_i(\xi), &\qquad \varphi''_{i,+}(x) &:= \lim_{\xi \searrow x} \varphi''_i(\xi) &\qquad \forall x \in [a_i^0, a_i^{m_i}).
\end{aligned}$$

exist in $\mathbb{R} \cup \{-\infty, \infty\}$.

Proof. The existence of the limits for $x \in (a_i^k, a_i^{k+1})$ and of the limits of φ''_i at the a_i^k is guaranteed by (A8). We only have to show the existence of the limits of φ'_i for $x = a_i^k$. First we note that φ'_i is monotone. Furthermore, it is bounded on each interval (a_i^{k-1}, a_i^k) with $k < m_i$ since φ_i cannot be convex on (a_i^{k-1}, a_i^{k+1}) otherwise. Thus $\lim_{\xi \nearrow a_i^k} \varphi'_i(\xi)$ exists and is finite for $k < m_i$ and either finite or ∞ for $k = m_i$. Limits from above can be shown analogously. □

Note that both one-sided derivatives coincide for $x \in (a_i^k, a_i^{k+1})$. For $x = a_i^k$ with $0 < k < m_i$ either the limits $\varphi'_{i,-}(x)$ and $\varphi'_{i,+}(x)$ or $\varphi''_{i,-}(x)$ and $\varphi''_{i,+}(x)$ do not coincide due to the maximality of the intervals. While in the latter case a one-sided second derivative makes sense this is no longer true in the former one since the subdifferential $\partial\varphi_i$ is set-valued in this case. In view of this fact and Lemma 4.2 we can now give linearizations of f_i in the following way:

Lemma 4.3. *Each f_i is single-valued, monotone and Lipschitz continuous. An element $\partial f_i(x)$ of the generalized Jacobian in the sense of Clarke $\partial_C f_i(x)$ is given by*

$$\partial f_i(x) = \begin{cases} 0 & \text{if } \partial\varphi_i(f_i(x)) \text{ is set-valued,} \\ (A_{ii} + \varphi_i''(f_i(x)))^{-1} & \text{else.} \end{cases} \tag{4.22}$$

For φ_i'' we use either the one-sided derivative $\varphi_{i,-}''$ from below or $\varphi_{i,+}''$ from above. Consequently $(A_{ii} + \varphi_i''(f_i(x)))^{-1}$ is set to zero if at least one one-sided derivative is infinite.

Proof. Single-valuedness, monotonicity and Lipschitz continuity follow directly from application of Proposition 4.1 to $\frac{1}{2}A_{ii}(\cdot)^2 + \varphi_i(\cdot)$. For the derivative we first note that the preimages $U_i^k = f_i^{-1}((a_i^k, a_i^{k+1}))$ are nonempty open intervals by continuity of f_i and strict monotonicity of the maximal monotone operator f_i^{-1}. Hence f_i is differentiable on U_i^k with

$$f_i'(x) = \big((f_i^{-1})'(f_i(x))\big)^{-1} = (A_{ii} + \varphi_i''(f_i(x)))^{-1} \in \mathbb{R} \qquad \forall x \in U_i^k.$$

Now we consider the preimages

$$f_i^{-1}(a_i^k) = A_{ii}a_i^k + \partial\varphi_i(a_i^k)$$

of a_i^k. If $\partial\varphi_i(a_i^k)$ is set-valued then $f_i^{-1}(a_i^k)$ is a non-trivial interval on which f_i is constant. Hence we have $0 = f_i'$ in the interior and $0 \in \partial f_i$ on the closure of the interval. If $\partial\varphi_i(a_i^k)$ is single-valued then $f_i^{-1}(a_i^k)$ is a single point. Since $\mathbb{R}$ decomposes into the previously mentioned preimages, elements of the generalized Jacobian at the finite number of isolated points $f_i^{-1}(a_i^k)$ can be given as limits of the derivatives in the sets U_i^k which are zero if φ_i'' is infinite. □

If we want to derive generalized linearizations of $\mathcal{F}_{GS}$ and $\mathcal{F}_J$ from derivatives of f_i, so-called truncated matrices appear naturally. In order to ease the handling of such matrices we introduce the following notation for truncated matrices and vectors. Some useful properties of such matrices and of the Moore–Penrose pseudoinverse

$$M^+ := \lim_{\epsilon\to 0}(M^TM + \epsilon I)^{-1}M^T \tag{4.23}$$

are collected in Section A.6 of the appendix.

Definition 4.2. *Let $\mathcal{I}, \mathcal{J} \subset \mathbb{N}$ be index sets, $x \in \mathbb{R}^n$ a vector, and $M \in \mathbb{R}^{m,n}$ a matrix. Then define the truncated matrix $M_{\mathcal{I},\mathcal{J}} \in \mathbb{R}^{m,n}$ and the truncated vector $x_{\mathcal{I}} \in \mathbb{R}^n$ by*

$$(M_{\mathcal{I}\mathcal{J}})_{ij} := \begin{cases} M_{ij} & \text{if } i \in \mathcal{I} \text{ and } j \in \mathcal{J}, \\ 0 & \text{else,} \end{cases} \qquad (x_{\mathcal{I}})_i := \begin{cases} x_i & \text{if } i \in \mathcal{I}, \\ 0 & \text{else.} \end{cases}$$

Furthermore, define the abbreviation $M_{\mathcal{I}} := M_{\mathcal{I},\mathcal{I}}$.

Having derivatives for f_i we are ready to construct linearizations of $\mathcal{F}_{GS}$ and $\mathcal{F}_J$. To this end let $x \in \mathbb{R}^n$ and $y = \mathcal{F}_{GS}(x)$. Starting from (4.12) we get

$$\mathcal{F}_{GS}(x) = (D + \partial\varphi)^{-1}\Big(b - (R+L)x - L\mathcal{F}_{GS}(x)\Big) - x.$$

Assuming a chain rule we can now define a generalized linearization of $\mathcal{F}_{GS}$ by

$$\partial\mathcal{F}_{GS}(x) := \partial\Big((D + \partial\varphi)^{-1}\Big)\Big(b - (R+L)x - L\mathcal{F}_{GS}(x)\Big)\Big(-(R+L) - L\partial\mathcal{F}_{GS}(x)\Big) - I.$$

Since the i-th component of $(D + \partial\varphi)^{-1}$ is just f_i we can plug in the derivatives of f_i at $r_i = (b - (R+L)x - L\mathcal{F}_{GS}(x))_i$ to get

$$\partial\mathcal{F}_{GS}(x) = \operatorname{diag}\big(\partial f_i(r_i)\big)\Big(-(R+L) - L\partial\mathcal{F}_{GS}(x)\Big)_{\mathcal{I}'(x+y)} - I$$

Now $f_i(r_i) = x_i + y_i$ and the representation of the Moore–Penrose pseudoinverse in Lemma A.5 yield

$$\partial\mathcal{F}_{GS}(x) = \Big(D + \varphi''(x+y)\Big)^{+}_{\mathcal{I}'(x+y)}\Big(-(R+L) - L\partial\mathcal{F}_{GS}(x)\Big) - I$$

with the index set of inactive components given by

$$\mathcal{I}(v) := \{i : \partial\varphi_i(v_i) \text{ is single-valued}\}, \tag{4.24}$$

$$\mathcal{I}'(v) := \big\{i \in \mathcal{I}(v) : \max\{\varphi''_{i,-}(v_i), \varphi''_{i,+}(v_i)\} < \infty\big\}. \tag{4.25}$$

While multiplication from the left by $I_{\mathbb{N}\setminus\mathcal{I}'(x+y)}$ implies

$$\partial\mathcal{F}_{GS}(x)_{\mathbb{N}\setminus\mathcal{I}'(x+y),\mathbb{N}} = I_{\mathbb{N}\setminus\mathcal{I}'(x+y)} \tag{4.26}$$

multiplication by $D + \varphi''(x+y)_{\mathcal{I}'(x+y)}$ provides

$$\big(D + \varphi''(x+y)\big)_{\mathcal{I}'(x+y)}\big(\partial\mathcal{F}_{GS}(x) + I\big) = \Big(-R - L\big(\partial\mathcal{F}_{GS}(x) + I\big)\Big)_{\mathcal{I}'(x+y),\mathbb{N}}$$

which is equivalent to

$$\big(D + L + \varphi''(x+y)\big)_{\mathcal{I}'(x+y),\mathbb{N}}\big(\partial\mathcal{F}_{GS}(x) + I\big) = -R_{\mathcal{I}'(x+y)}$$

and hence

$$\begin{aligned}\partial\mathcal{F}_{GS}(x)_{\mathcal{I}'(x+y),\mathbb{N}} &= -\Big(D + L + \varphi''(x+y)\Big)^{+}_{\mathcal{I}'(x+y),\mathbb{N}} R_{\mathcal{I}'(x+y)} - I_{\mathcal{I}'(x+y)} \\ &= -\Big(D + L + \varphi''(x+y)\Big)^{+}_{\mathcal{I}'(x+y)} R_{\mathcal{I}'(x+y)} - I_{\mathcal{I}'(x+y)}.\end{aligned}$$

Together with (4.26) we get

$$\partial\mathcal{F}_{GS}(x) = -\Big(D + L + \varphi''(x+y)\Big)^{+}_{\mathcal{I}'(x+y)} R_{\mathcal{I}'(x+y)} - I. \tag{4.27}$$

Using the same arguments we can derive

$$\partial\mathcal{F}_{J}(x) := -\Big(D + \varphi''(x+y)\Big)^{+}_{\mathcal{I}'(x+y)} (L+R)_{\mathcal{I}'(x+y)} - I. \tag{4.28}$$

4.2.4 Nonsmooth Newton Methods for Nonlinear Smoothers

Now we want to apply Newton-type methods

$$u^{\nu+1} = u^\nu - H(u^\nu)^{-1}\mathcal{F}(u^\nu) \tag{4.29}$$

for the iterative solution of (4.21). Unfortunately the linearizations $\partial\mathcal{F}_{GS}$ and $\partial\mathcal{F}_J$ incorporate inverse matrices and are not positive semidefinite although the original problem is convex. To overcome this drawback we now present a reformulation in terms of minimization algorithms for the energy J. This has also the advantage that it provides a natural way to globalize the methods.

Lemma 4.4. *Assume that J_0 and φ satisfy (A6) and (A8), respectively. Then the derivative of J restricted to the smooth components given by*

$$\nabla J(y)_{\mathcal{I}'(y)} := (\nabla J_0 + \varphi')(y)_{\mathcal{I}'(y)} = (Ay - b + \varphi'(y))_{\mathcal{I}'(y)}$$

satisfies

$$\begin{aligned}\nabla J(x + \mathcal{F}_{GS}(x))_{\mathcal{I}'(x+\mathcal{F}_{GS}(x))} &= R_{\mathcal{I}'(x+\mathcal{F}_{GS}(x))}\mathcal{F}_{GS}(x),\\ \nabla J(x + \mathcal{F}_J(x))_{\mathcal{I}'(x+\mathcal{F}_J(x))} &= (L+R)_{\mathcal{I}'(x+\mathcal{F}_J(x))}\mathcal{F}_J(x).\end{aligned}$$

Proof. (4.12) and (4.13) yield

$$\begin{aligned}(D + L + \partial\varphi)(x + \mathcal{F}_{GS}(x)) &\ni b - Rx,\\ (D + \partial\varphi)(x + \mathcal{F}_J(x)) &\ni b - (L+R)x.\end{aligned}$$

Adding $R\mathcal{F}_{GS}(x) - b$ and $(L+R)\mathcal{F}_J(x) - b$ to these equations we get

$$\begin{aligned}(\nabla J_0 + \partial\varphi)(x + \mathcal{F}_{GS}(x)) &\ni R\mathcal{F}_{GS}(x),\\ (\nabla J_0 + \partial\varphi)(x + \mathcal{F}_J(x)) &\ni (L+R)\mathcal{F}_J(x).\end{aligned}$$

Restriction to the inactive sets provides the assertion. □

Theorem 4.4. *Let $\mathcal{F} = \mathcal{F}_{GS}$ or $\mathcal{F} = \mathcal{F}_J$. If $H(u^\nu) = \partial\mathcal{F}(u^\nu)$ is used in a nonsmooth Newton step* (4.29) *the resulting iteration can be equivalently rewritten as the following two-stage method*

$$u^{\nu+\frac{1}{2}} = u^\nu + \mathcal{F}(u^\nu), \tag{4.30}$$

$$u^{\nu+1} = u^{\nu+\frac{1}{2}} + \mathcal{C}(u^{\nu+\frac{1}{2}}), \tag{4.31}$$

with the linear correction operator

$$\mathcal{C}(v) := -\Big(A + \varphi''(v)\Big)^{+}_{\mathcal{I}'(v)} \nabla J(v)_{\mathcal{I}'(v)}. \tag{4.32}$$

Proof. We give the proof for $\mathcal{F} = \mathcal{F}_{GS}$. For simplicity define $v = u^{\nu+\frac{1}{2}}$ by (4.30). Inserting $\partial\mathcal{F}_{GS}$ in the equation for the Newton correction we get

$$\left[\Big(D + L + \varphi''(v)\Big)^{+}_{\mathcal{I}'(v)} R_{\mathcal{I}'(v)} + I\right] (u^{\nu+1} - u^{\nu}) = \mathcal{F}(u^{\nu})$$

and hence

$$(u^{\nu+1} - v)_{\mathbb{N}\setminus\mathcal{I}'(v)} = (u^{\nu+1} - u^{\nu})_{\mathbb{N}\setminus\mathcal{I}'(v)} - \mathcal{F}(u^{\nu})_{\mathbb{N}\setminus\mathcal{I}'(v)} = 0 = \mathcal{C}(v)_{\mathbb{N}\setminus\mathcal{I}'(v)}.$$

Furthermore, we have

$$R_{\mathcal{I}'(v)}(u^{\nu+1} - u^{\nu}) = \Big(D + L + \varphi''(v)\Big)_{\mathcal{I}'(v)} (v - u^{\nu+1}).$$

Adding $R_{\mathcal{I}'(v)}(v - u^{\nu+1})$ and using Lemma 4.4 we get

$$\nabla J(v)_{\mathcal{I}'(v)} = R_{\mathcal{I}'(v)}\mathcal{F}(u^{\nu}) = \Big(A + \varphi''(v)\Big)_{\mathcal{I}'(v)} (v - u^{\nu+1})$$

and finally

$$(u^{\nu+1} - v)_{\mathcal{I}'(v)} = -\Big(A + \varphi''(v)\Big)^{+}_{\mathcal{I}'(v)} \nabla J(v)_{\mathcal{I}'(v)} = \mathcal{C}(v)_{\mathcal{I}'(v)}.$$

Thus we have shown $u^{\nu+1} - v = \mathcal{C}(v)$.

The proof for $\mathcal{F} = \mathcal{F}_J$ is obtained by replacing L by 0 and R by $L + R$, respectively. □

The reformulation of the nonsmooth Newton method in Theorem 4.4 has the advantage that the linear subproblem to be solved in each iteration step has a much nicer structure. The matrix $(A + \varphi''(v))_{\mathcal{I}'(v)}$ is symmetric and, due to convexity of J and truncation, positive semidefinite. Furthermore, its kernel is known since it is spanned by the Euclidean unit vectors e_i with $i \notin \mathcal{I}'(v)$. By symmetry the matrix is even positive definite on its range which is just the reduced subspace

$$V_{\mathcal{I}'(v)} = \operatorname{span}\{e_i : i \in \mathcal{I}'(v)\} \tag{4.33}$$

obtained by omitting the rows and columns with indices $i \notin \mathcal{I}'(v)$. Thus common iterative methods like the preconditioned conjugate gradient method or multigrid methods can be applied for the iterative solution of (4.32).

Remark 4.1. *While $\nabla J(v)_{\mathcal{I}'(v)}$ is the gradient of J reduced to the subspace $V_{\mathcal{I}'(v)}$ where it is differentiable, the matrix $(A + \varphi''(v))_{\mathcal{I}'(v)}$ represents a linearization of $\nabla J(v)_{\mathcal{I}'(v)}$ and thus a reduced Hessian of J on this subspace. Hence* (4.30) *and* (4.31) *can also be regarded as subsequent application of the nonlinear smoother on the whole space and a Newton step on the reduced space $V_{\mathcal{I}'(u^{\nu+\frac{1}{2}})}$.*

Remark 4.2. *For twice continuously differentiable φ the matrix in* (4.32) *reduces to the Hessian of J at $u^{\nu+\frac{1}{2}}$. In case of a nonquadratic J_0 the application of a classical Newton step also leads to a formulation similar to* (4.30) *and* (4.31)*. However, in this case the matrix in* (4.32) *does not contain the Hessian of J. It contains a matrix consisting of the second partial derivatives of J at u^ν above and at $u^{\nu+\frac{1}{2}}$ on and below the diagonal.*

In light of Remarks 4.1 and 4.2 we can generalize the method to other smooth nonlinearities. Assume that (A3), (A8), and (A4) are satisfied. Then we define the reduced gradient and the reduced Hessian of J at v on the subspace $V_{\mathcal{J}}$ for an index set $\mathcal{J} \subset \mathcal{I}'(v)$ by

$$\nabla J(v)_{\mathcal{J}} := \Big(\nabla J_0(v) + \varphi'(v)\Big)_{\mathcal{J}}, \qquad \partial^2 J(v)_{\mathcal{J}} := \Big(\partial^2 J_0(v) + \varphi''(v)\Big)_{\mathcal{J}}. \tag{4.34}$$

Now (4.32) can be generalized by defining

$$\mathcal{C}(v) := -\Big(\partial^2 J(v)_{\mathcal{I}'(v)}\Big)^+ \nabla J(v)_{\mathcal{I}'(v)}. \tag{4.35}$$

Note that this is not equivalent to the application of classical Newton method for twice continuously differentiable J (see Remark 4.2) since we use the more recent information of $u^{\nu+\frac{1}{2}}$ even on and below the diagonal.

Even though the linearization of ∇J in (4.34) is restricted to components where φ is locally smooth the second derivatives φ_i'' might get very large leading to arbitrarily ill-conditioned linear systems. This effect appears additionally to a possible ill-conditioning of $\partial^2 J_0(v)$ resulting from a discretized differential operator. If the linear systems are solved iteratively this will in general lead to a considerable slowdown even if multigrid methods are applied. Therefore, we restrict the linearization further to

$$\mathcal{I}''(v) := \{i \in \mathcal{I}'(v) : \varphi_i''(v_i) < (C_\varphi)_{i,i}\}. \tag{4.36}$$

where $C_\varphi \in \mathbb{R}^{n,n}$ is a positive definite diagonal matrix.

Remark 4.3. *Replacing $\mathcal{I}'(u^{\nu+\frac{1}{2}})$ by some $\overline{\mathcal{I}'} \subset \mathcal{I}'(u^{\nu+\frac{1}{2}})$ leads to truncated linearizations $\overline{\partial}\mathcal{F}_{GS}$ and $\overline{\partial}\mathcal{F}_J$ defined analogously to* (4.27) *and* (4.28)*, respectively. Theorem 4.4 remains true for $H(u^k) = \overline{\partial}\mathcal{F}(u^k)$ if $\mathcal{I}'(u^{\nu+\frac{1}{2}})$ is replaced by the smaller index set $\overline{\mathcal{I}'}(u^{\nu+\frac{1}{2}})$.*

In general we will truncate large second derivatives of φ by using large constants C_φ. In light of (4.27) and (4.28) such derivatives would result in small derivatives of $\mathcal{F}_{GS}$ and $\mathcal{F}_J$. Hence the additional truncation essentially means to set small derivatives to zero for these operators.

4.2.5 Convergence Analysis

Now we consider the convergence analysis of the previously presented algorithms. These algorithms have the considerable advantage that the reformulation in terms of a smoothing operator allows to apply a Newton-type approach even if this is not directly possible for the minimization problem. Unfortunately, the algorithms depend heavily on the selection of the basis. If we are faced with a sequence of problems resulting from discretization of a partial differential equation we can in general not expect grid independent convergence as the following example shows.

Example 4.2. *Let $\mathcal{S}^h \subset H^1_0(-1,1)$ be the space of linear finite elements on a uniform grid with mesh size $h = 1/(n+1)$ and nodes x_i on $(-1,1)$. Discretizing the minimization problem*

$$u = \operatorname*{arg\,min}_{v \in K} \int_0^1 (v'(x))^2 dx \tag{4.37}$$

with $K = \{v \in H^1_0(-1,1) : v \geq -1 \,\text{a.e.}\}$ by considering it in the discrete convex set

$$K^h = \{v \in \mathcal{S}^h : v(x_i) \geq -1, i = 1, \dots, n\}$$

leads to a discrete problem of the form (4.1) *satisfying (A6) and (A8) where A is the usual stiffness matrix. Assume that the nodes x_i are ordered such that $i < j$ if $|x_i| < |x_j|$. Now consider the algorithm in Theorem 4.4 with $\mathcal{F} = \mathcal{F}_{GS}$ or $\mathcal{F} = \mathcal{F}_J$, an arbitrary correction $\mathcal{C}(v) \in V_{\mathcal{I}'(v)}$, the initial value $u^0_h = -1$ and hence $\mathcal{I}'(u^0_h) = \emptyset$. Since the correction $\mathcal{C}(v)$ is restricted to $V_{\mathcal{I}'(v)}$ only the application of $\mathcal{F}$ can enlarge $\mathcal{I}'(u^\nu)$. But because u^ν is constant on the set* $\operatorname{co}\{x_i : i \notin \mathcal{I}'(u^\nu)\}$ *only the leftmost and rightmost nodes of this set might enter $\mathcal{I}'(u^{\nu+1})$. Hence we know $u^\nu = -1$ on the set $(-1+(\nu+1)h, 1-(\nu+1)h)$ and thus*

$$\|u^\nu - u^*\| = \|u^\nu\| \geq 2 - 2\frac{\nu+1}{n+1}.$$

This example shows that finding the exact inactive set $\mathcal{I}'(u^*)$ might take $\mathcal{O}(n)$ iteration steps. Furthermore, the linear subproblem as proposed in Theorem 4.4 already provides the exact solution once the correct inactive set $\mathcal{I}'(u^*)$ is detected.

In light of this example it will be hopeless to look for mesh-independent convergence results. Even local convergence results might only cover the region of finite termination. Having this in mind we now aim at global convergence results incorporating inexact evaluation of the linear subproblems. With such a result we can still hope that the speed of the linear subproblem solver dominates at least asymptotically or for reasonable initial iterates.

Theorem 4.1 gives a general convergence result for the class of algorithms we consider. However, the corrections $\mathcal{C}(u^{\nu+\frac{1}{2}})$ need not be monotone. It is even possible that they leave $\operatorname{dom}(J)$ and hence produce infinite energy. Even for smooth problems, where these corrections reduce to classical Newton corrections, monotonicity is not guaranteed. A common remedy in classical Newton methods is to apply damping according to the energy. While it is crucial to select appropriate damping strategies in the smooth case to preserve properties of the local Newton convergence (see [43]) the situation is not so clear in the presented nonsmooth case.

Since a local mesh-independent convergence theory is not at hand we need not preserve properties of local Newton methods. Furthermore, we only want to apply damping to the reduced Newton-type correction $\mathcal{C}(u^{\nu+\frac{1}{2}})$ and not to the whole correction $\mathcal{F}(u^\nu) + \mathcal{C}(u^\nu + \mathcal{F}(u^\nu))$ to allow the application of Theorem 4.1. Thus we stick to a damping strategy based on simple energy minimization. A first attempt would be to replace $\mathcal{C}(u^{\nu+\frac{1}{2}})$ by

$$\rho\left(J, u^{\nu+\frac{1}{2}}, \mathcal{C}(u^{\nu+\frac{1}{2}})\right) \mathcal{C}(u^{\nu+\frac{1}{2}}). \tag{4.38}$$

With this strategy the damping parameters are a priori bounded from above by

$$\min\left\{ \min_{i\in\mathcal{I}_+} \frac{\max(\operatorname{dom}\varphi_i) - u_i^{\nu+\frac{1}{2}}}{\mathcal{C}(u^{\nu+\frac{1}{2}})}, \min_{i\in\mathcal{I}_-} \frac{\min(\operatorname{dom}\varphi_i) - u_i^{\nu+\frac{1}{2}}}{\mathcal{C}(u^{\nu+\frac{1}{2}})} \right\} \tag{4.39}$$

with

$$\mathcal{I}_+ = \{i : \mathcal{C}(u^{\nu+\frac{1}{2}})_i > 0\}, \qquad \mathcal{I}_- = \{i : \mathcal{C}(u^{\nu+\frac{1}{2}})_i < 0\}.$$

This might enforce arbitrarily small damping parameters leading to slow convergence if $u^{\nu+\frac{1}{2}}$ almost touches $\operatorname{dom}(J)$ even at a single component.

To avoid this problem we introduce a projection before damping is applied. For $x \in \operatorname{dom}\varphi$ the orthogonal projection of d into $(\operatorname{dom}\varphi) - x$ is given by

$$(P_{(\operatorname{dom}\varphi)-x} d)_i := \max\left\{\min\left\{d_i, \max\left((\operatorname{dom}\varphi_i) - x_i\right)\right\}, \min\left((\operatorname{dom}\varphi_i) - x_i\right)\right\}.$$

It does always satisfy $x + P_{(\operatorname{dom}\varphi)-x} d \in \operatorname{dom}\varphi$. Now we are ready to state the modified algorithms. The following global convergence results are a direct consequence of Theorem 4.1.

Corollary 4.3. *Assume that (A3), (A8), and (A4) are satisfied and that $u^0 \in \operatorname{dom}\varphi$. Then the sequence u^ν generated by*

$$u^{\nu+\frac{1}{2}} = u^\nu + \mathcal{F}_{GS}(u^\nu), \tag{4.40}$$

$$u^{\nu+1} = u^{\nu+\frac{1}{2}} + \rho^\nu P^\nu \left(\mathcal{C}(u^{\nu+\frac{1}{2}}) + \epsilon^\nu\right), \tag{4.41}$$

with the truncated linear correction

$$\mathcal{C}(v) := -\left(\partial^2 J(v)_{\mathcal{I}''(v)}\right)^+ \nabla J(v)_{\mathcal{I}''(v)},$$

the projection $P^\nu = P_{(\operatorname{dom}\varphi)-u^{\nu+\frac{1}{2}}}$ and damping parameter

$$\rho^\nu \in \rho_<\left(J, u^{\nu+\frac{1}{2}}, P^\nu\left(\mathcal{C}(u^{\nu+\frac{1}{2}}) + \epsilon^\nu\right), 0\right)$$

converges to the unique solution of (4.1) *for every ϵ^ν. The same is true if $\mathcal{F}_{GS}$ is replaced by its inexact version $\overline{\mathcal{F}}_{GS}$ as introduced in Theorem 4.2.*

Corollary 4.4. *Assume that (A3), (A8), and (A4) are satisfied and that* $u^0 \in \operatorname{dom}\varphi$*. Then the sequence* u^ν *generated by*

$$u^{\nu+\frac{1}{2}} = u^\nu + \rho_J^\nu \mathcal{F}_J(u^\nu), \tag{4.42}$$

$$u^{\nu+1} = u^{\nu+\frac{1}{2}} + \rho^\nu P^\nu \left(\mathcal{C}(u^{\nu+\frac{1}{2}}) + \epsilon^\nu\right), \tag{4.43}$$

with the truncated linear correction

$$\mathcal{C}(v) := -\left(\partial^2 J(v)_{\mathcal{I}''(v)}\right)^+ \nabla J(v)_{\mathcal{I}''(v)},$$

the projection $P^\nu = P_{(\operatorname{dom}\varphi)-u^{\nu+\frac{1}{2}}}$ *and damping parameters*

$$\rho_J^\nu \in \rho_<\left(J, u^\nu, \mathcal{F}_J(u^\nu), \frac{1}{n}\right), \qquad \rho^\nu \in \rho_<\left(J, u^{\nu+\frac{1}{2}}, P^\nu\left(\mathcal{C}(u^{\nu+\frac{1}{2}}) + \epsilon^\nu\right), 0\right)$$

converges to the unique solution of (4.9) *for every* ϵ^ν*. The same is true if* $\mathcal{F}_J$ *is replaced by its inexact version* $\overline{\mathcal{F}}_J$ *as introduced in Theorem 4.3.*

4.2.6 Relation to Primal–Dual Active Set Methods

In order to relate the presented algorithm to other active set type methods we assume the case of a quadratic obstacle problem presented in Example 4.1 with $\underline{\psi} = -\infty$.

A class of well-known algorithms for this problem are primal and primal–dual active set methods. The primal active set method for obstacle problems was introduced by Hoppe [65]. The closely related primal–dual active set method was introduced in Bergounioux et al. [14]. Hintermüller et al. [64] showed that this method can be regarded as semismooth Newton method. Since the primal method coincides with the primal–dual method after the first iteration step [64] we only consider the latter here using the abbreviation PDAS. It is based on the primal–dual formulation for the obstacle problem given by

$$\begin{gathered} Au^* + \lambda^* = b, \\ u^* \le \overline{\psi}, \qquad \lambda^* \ge 0, \qquad \lambda^*(u^* - \overline{\psi}) = 0 \end{gathered}$$

where $\lambda^* \in \mathbb{R}^n$ is the Lagrangian multiplier for the inequality constraints. For given λ^0,u^0 and a fixed constant $c > 0$ the primal–dual active set method reads as follows

$$\mathcal{A}_\nu := \{i : \lambda_i^\nu + c(u_i^\nu - \overline{\psi}_i) > 0\}, \tag{4.44}$$

$$(u^{\nu+1}, \lambda^{\nu+1}) \in \mathbb{R}^n \times \mathbb{R}^n : \qquad \begin{array}{rcll} Au^{\nu+1} + \lambda^{\nu+1} & = & b, & \\ u_i^{\nu+1} & = & \overline{\psi}_i & \text{for } i \in \mathcal{A}_\nu, \\ \lambda_i^{\nu+1} & = & 0 & \text{for } i \notin \mathcal{A}_\nu. \end{array} \tag{4.45}$$

After the first iteration this algorithm behaves very similar to the presented algorithm based on the nonlinear Jacobi iteration.

Theorem 4.5. *Let $\nu \geq 1$. Then the PDAS method takes the form*

$$u^{\nu+\frac{1}{2}} = u^\nu + \mathcal{F}_J(u^\nu), \tag{4.46}$$

$$u^{\nu+1} = u^{\nu+\frac{1}{2}} + \mathcal{C}(u^{\nu+\frac{1}{2}}, u^\nu), \tag{4.47}$$

with the linear correction

$$\mathcal{C}(v,w) = -\Big(A + \varphi''(v)\Big)^+_{\mathcal{J}(v,w)} \nabla J(v)_{\mathcal{J}(v,w)} \tag{4.48}$$

and the inactive set

$$\mathcal{J}(v,w) := \mathcal{I}'(v) \cup \left\{ i : \overline{\psi}_i = w_i + a_{ii}^{-1}(b - Aw)_i \right\}.$$

Furthermore, the modified PDAS method with $\mathcal{A}_\nu$ replaced by

$$\widetilde{\mathcal{A}}_\nu = \{ i : \lambda_i^\nu + c(u_i^\nu - \overline{\psi}_i) \geq 0 \}$$

coincides with the nonsmooth Newton method in Theorem 4.4 for $\mathcal{F} = \mathcal{F}_J$.

Proof. First we note that the nonlinear Jacobi step can be represented by

$$\mathcal{F}_J(v)_i = \min\left\{ a_{ii}^{-1}(b - Av)_i, \overline{\psi}_i - v \right\}.$$

By the assumption $\nu \geq 1$ we have $\lambda^\nu = b - Au^\nu$ and for all i either $\lambda_i = (b - Au^\nu)$ or $\overline{\psi}_i - u_i^\nu$ are zero. Hence

$$\begin{aligned}
i \in \mathcal{A}_\nu &\Leftrightarrow \Big[(b - Au^\nu)_i > 0 \text{ or } \overline{\psi}_i - u_i^\nu < 0 \Big] \\
&\Leftrightarrow \overline{\psi}_i - u_i^\nu < (b - Au^\nu)_i \\
&\Leftrightarrow \Big[\mathcal{F}_J(u^\nu)_i = \overline{\psi}_i - u_i^\nu \text{ and } a_{ii}^{-1}(b - Au^\nu)_i \neq \overline{\psi}_i - u_i^\nu \Big] \\
&\Leftrightarrow i \notin \mathcal{I}_\nu := \mathcal{J}(u^{\nu+\frac{1}{2}}, u^\nu).
\end{aligned}$$

Now assume that $u^{\nu+1}$ is defined by equations (4.46)–(4.47). Then we have $u_i^{\nu+1} = u_i^{\nu+\frac{1}{2}} = \overline{\psi}_i$ for $i \in \mathcal{A}_\nu$. Since $u^{\nu+\frac{1}{2}} \leq \overline{\psi}$ the left-sided first and second derivatives of φ_i at $u_i^{\nu+\frac{1}{2}}$ are zero. Inserting this in (4.47) and multiplying by $A_{\mathcal{I}_\nu}$ we get

$$\begin{aligned}
A_{\mathcal{I}_\nu} u^{\nu+1} &= A_{\mathcal{I}_\nu} u^{\nu+\frac{1}{2}} + (b - Au^{\nu+\frac{1}{2}})_{\mathcal{I}_\nu} \\
&= \Big(b - (A - A_{\mathbb{N},\mathcal{I}_\nu}) u^{\nu+\frac{1}{2}} \Big)_{\mathcal{I}_\nu} \\
&= b_{\mathcal{I}_\nu} - A_{\mathcal{I}_\nu,\mathcal{A}_\nu} u^{\nu+\frac{1}{2}} = b_{\mathcal{I}_\nu} - A_{\mathcal{I}_\nu,\mathcal{A}_\nu} u^{\nu+1}
\end{aligned}$$

and hence $(Au^{\nu+1} - b)_{\mathcal{I}_\nu} = 0$. Setting $\lambda^{\nu+1} := b - Au^{\nu+1}$ we get the equivalence to the primal–dual active set method.

For the modified method we can similarly get $i \in \widetilde{\mathcal{A}}_\nu \Leftrightarrow i \notin \widetilde{\mathcal{I}}_\nu := \mathcal{I}'(u^{\nu+\frac{1}{2}})$ and then use the same proof with $\mathcal{A}_\nu$ and $\mathcal{I}_\nu$ replaced by $\widetilde{\mathcal{A}}_\nu$ and $\widetilde{\mathcal{I}}_\nu$, respectively. □

After the first step the only difference between the PDAS method and the nonsmooth Newton method for $\mathcal{F}_J$ in Theorem 4.4 is the slightly larger inactive set $\mathcal{I}_\nu$ compared to $\widetilde{\mathcal{I}}_\nu$ which also incorporates $\left\{i : \overline{\psi}_i = w_i + a_{ii}^{-1}(b - Aw)_i\right\}$. This is the set of all indices for which the solution of the one dimensional constrained quadratic minimization problem in the nonlinear Jacobi step coincides with the solution of the corresponding unconstrained quadratic problem. This degenerate case will in general only rarely happen.

Thus Theorem 4.5 shows that for the case of an unilaterally constrained obstacle problem the PDAS method and the method in Theorem 4.4 are essentially the same. Since the presented formulation also incorporates the Gauß–Seidel smoother and other smooth and nonsmooth nonlinearities J_0 and φ it can be regarded as a generalization of the PDAS method.

Currently there are two types of convergence results for the PDAS method. The first one provides global convergence if A is an M-matrix and if the linear systems are solved exactly [64, 65]. Since problems arising from adaptive finite element discretizations do not lead to the M-matrix property in general, this result is rather unsatisfactory. Furthermore, it can be seen in numerical experiments that global convergence is in general not preserved for inexact evaluation of the linear systems using e.g. one linear multigrid step.

The second convergence result relies on a reformulation of the algorithm as semismooth Newton method and the so-called slanting differentiability of the operator appearing in this formulation [64]. While this does only provide local convergence it offers the possibility to solve the linear systems inexactly up to a tolerance required by the surrounding Newton type method. Although the method and this convergence result can be extended to function spaces [64, 67] e.g. for operators $A : L^2 \to L^2$, this is in general not possible if A is a differential operator because the Lagrangian multipliers are only measures in this case. Since the convergence proof for the discrete case does not give any knowledge on the domain of convergence it is in general not excluded that u^0, λ^0 being in this domain requires

$$\mathcal{A}_0 = \{i : u_i^* = \overline{\psi}_i\}.$$

However, in this case the problem reduces to a linear problem and one step convergence is obvious.

Note that the presented formulation also relies on a Newton type method. However, the convergence analysis is based on energy descent. This allows to globalize the method using monotonicity arguments for the natural energy J of the problem at hand. The construction of artificial merit functions is not necessary. This global convergence theory is still valid in case of inexact evaluation.

4.3 Truncated Nonsmooth Newton Multigrid

It is crucial for an efficient overall algorithm to use a fast method for the inexact solution of the linear subproblems. In this section we introduce a multigrid method for the linear subproblems and discuss the relation of the obtained overall method to other nonlinear multigrid methods for special cases of the minimization problem (4.1).

4.3.1 Multigrid Solution of Truncated Linear Problems

Multigrid methods are known to be one of the fastest approaches to solve linear systems resulting from discretized partial differential equations. Originally these methods were based on a sequence of discrete problems on a hierarchy of grids [20, 63]. Additionally to the grid for the current problem (called fine grid) there are also coarser grids and corresponding discretizations. One step of a multigrid method applies several so-called smoothing steps to the problem on each level. These smoothing steps are often applications of simple relaxation methods (like the Gauß–Seidel or the Jacobi method) called smoothers. While the smoother on the fine grid is supposed to reduce the high frequency parts of the algebraic error the low frequency parts can be resolved by relaxation on coarser grids.

For symmetric positive semidefinite systems one can also formulate multigrid methods as subspace correction methods for the associated minimization problems. The framework of subspace correction methods [111, 113] unifies multigrid and domain decomposition convergence theory and allows for mesh-independent convergence results without regularity assumptions on the solution of the underlying partial differential equation [23, 24].

Our interest is to introduce a multigrid method for truncated linear systems

$$A_{\mathcal{I}} u = b_{\mathcal{I}} \tag{4.49}$$

that appear in the evaluation of the correction $\mathcal{C}(u^{\nu+\frac{1}{2}})$ in the algorithms in Corollary 4.3 and Corollary 4.4 in the previous section. While solutions of this system are only unique in the components $i \in \mathcal{I}$, we originally aim at the computation of $u = A_{\mathcal{I}}^{+} b_{\mathcal{I}}$ and thus consider $u \in \{y \in \mathbb{R}^n : v_i = 0 \, \forall u \in \mathcal{I}\}$ only.

In order to introduce a multigrid method for this problem class we first recall the basic ingredients for a subspace correction formulation of a linear multigrid method for an untruncated problem

$$u \in V^j : \qquad a(u,v) = l(v) \qquad \forall v \in V^j \tag{4.50}$$

in a finite-dimensional space V^j with a symmetric positive define bilinear form a and $l \in (V^j)'$. Note that this can be written as $Ax = b$ (without truncation) where x, A, b represent u, a, l, respectively, with respect to a basis. Now let $V^0 \subset \cdots \subset V^j$ be a sequence of nested subspaces of V^j and $b_k : V^k \times V^k \to \mathbb{R}$ an approximation of a on V^k. Then one step of the successive subspace correction method for a given iterate u^ν

is defined by

```
$r^j = l - a(u^\nu, \cdot)$,
for $k = j, \dots, 0$ :
    solve $v^k \in V^k$ :    $b_k(v^k, v) = r^k(v) \qquad \forall v \in V^k$,
    if $k > 0$ :
        $r^{k-1} = r^k - a(v^k, \cdot)$,
$u^{\nu+1} = u^\nu + \sum_{k=0}^{j} v^k$
```

Now assume that a basis of V^k is given by

$$V^k = \operatorname{span}\{\lambda_1^k, \dots, \lambda_{n_k}^k\}, \qquad n_k = \dim V^k.$$

Then the natural embedding of V^k in V^{k+1} by the identity is represented by the transfer operators T_k satisfying

$$T_k \in \mathbb{R}^{n_{k+1}, n_k} : \qquad \lambda_i^k = \sum_{m=1}^{n_{k+1}} (T_k)_{mi} \lambda_m^{k+1} \qquad \forall i = 1, \dots, n_k$$

in the sense that they satisfy

$$v^{(k+1)} = T_k v^{(k)}$$

for two representations $v^{(k)} \in \mathbb{R}^{n_k}$ and $v^{(k+1)} \in \mathbb{R}^{n_{k+1}}$ of some $v \in V^k$ with respect to the bases in V^k and $V^{k+1} \supset V^k$, respectively. Note that for $k' > k$ the representation of the basis functions in V^k with respect to the basis in $V^{k'}$ is given by

$$\lambda_i^k = \sum_{m=1}^{n_{k'}} \left(T_{k'-1} \dots T_k\right)_{mi} \lambda_m^{k'}. \tag{4.51}$$

From now on we identify all vectors in the spaces $V^j, \dots, V^0$ with suitable representations in $\mathbb{R}^{n_j}, \dots, \mathbb{R}^{n_0}$ with respect to the given bases again. Then we can reformulate the multigrid method in algebraic form using the transfer operators:

```
$A_j = A$,
$r^j = b - A_j u^\nu$,
for $k = j, \dots, 0$ :
    $v^k = B_k^{-1} r^k$,
    if $k > 0$ :
        $r^{k-1} = T_{k-1}^T \left(r^k - A_k v^k\right)$,
        $A_{k-1} = T_{k-1}^T A_k T_{k-1}$,
$u^{\nu+1} = u^\nu + \sum_{k=0}^{j} \left(T_{j-1} \dots T_k\right) v^k$.
```

Here A_k and B_k denote the representations of a and b_k with respect to the basis in V^k.

Remark 4.4. *If the minimization problem results from a partial differential equation discretized in the first-order finite element space*

$$V^j = \mathcal{S}\Big(\mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_j)\Big)$$

on the leaf grid $\mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_j)$ of a grid hierarchy $(\mathcal{T}_0, \dots, \mathcal{T}_j)$, then by Corollary 3.1 a natural hierarchy of subspaces and corresponding bases is given by

$$V^k = \mathcal{S}\Big(\mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_k)\Big), \qquad \{\lambda_1^k, \dots, \lambda_{n_k}^k\} = \mathcal{B}\Big(\mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_k)\Big).$$

This hierarchy was used by Bramble et al. [22] for a parallel multigrid preconditioner. The resulting multigrid method is denoted "full multigrid". Note that this is in general not optimal in the sense that the computational effort is not in $O(n_j)$ but only in $O(n_j^2)$ for strongly local refinement. Yserentant [112] showed that optimal complexity can be achieved by restricting the corrections to suitable subspaces of V^k that do not build a nested hierarchy themselves.

If we want to use this method for truncated systems (4.49) with A replaced by $A_\mathcal{I}$ for varying $\mathcal{I}$ we have to introduce some small modifications. Since the matrix $A_\mathcal{I}$ is only positive semidefinite and thus not invertible, the same may happen for the restrictions $\widetilde{A}_k$. This will in general carry over to the matrices $\widetilde{B}_k$ if they are constructed in some way from $\widetilde{A}_k$, analogously to B_k.

For example the Gauß–Seidel smoother for $\widetilde{A}_k$ is given by $\widetilde{B}_k = \widetilde{D}_k + \widetilde{L}_k$ where $\widetilde{D}_k$ and $\widetilde{L}_k$ denote the diagonal and left part of $\widetilde{A}_k$, respectively. Consequently the inverse of $\widetilde{B}_k$ is replaced by the Moore–Penrose pseudoinverse $\widetilde{B}_k^+$ which coincides with the inverse if $\widetilde{B}_k$ is invertible. The application of $\widetilde{B}_k^+$ is equivalent to the subsequent scalar minimization in the i-th coordinate directions with the exception that the i-th scalar correction is set to zero if $(\widetilde{D}_k)_{ii} = 0$.

To guarantee that the iterates stay within the set $\{y \in \mathbb{R}^n : v_i = 0 \ \forall u \in \mathcal{I}\}$ we introduce a projection $I_\mathcal{I}$ of the sum of all corrections leading to the algorithm

```
$\widetilde{A}_j = A_\mathcal{I} = I_\mathcal{I} A I_\mathcal{I},$
$r^j = b_\mathcal{I} - \widetilde{A}_j u^\nu,$
for $k = j, \dots, 0$ :
    $v^k = \widetilde{B}_k^+ r^k,$
    if $k > 0$ :
        $r^{k-1} = T_{k-1}^T \left( r^k - \widetilde{A}_k v^k \right),$
        $\widetilde{A}_{k-1} = T_{k-1}^T \widetilde{A}_k T_{k-1},$
$u^{\nu+1} = u^\nu + \sum_{k=0}^{j} I_\mathcal{I} \left( T_{j-1} \dots T_k \right) v^k.$
```

The above method does not change if the transfer operator T_{j-1} is replaced by $I_{\mathcal{I}}T_{j-1}$. Hence it can also be interpreted in terms of the truncated subspaces

$$V^k_{\mathcal{I}} = P_{V^j_{\mathcal{I}}} V^k = \operatorname{span}\{P_{V^j_{\mathcal{I}}} \lambda^k_i : i = 1, \dots, n_k\} \tag{4.52}$$

where

$$P_{V^j_{\mathcal{I}}} : V^j \to V^j_{\mathcal{I}} = \operatorname{span}\{\lambda^j_i : i \in \mathcal{I}\}, \qquad P_{V^j_{\mathcal{I}}} \lambda^j_i = \begin{cases} \lambda^j_i & \text{if } i \in \mathcal{I}, \\ 0 & \text{else,} \end{cases}$$

is the orthogonal projection to $V^j_{\mathcal{I}}$ with respect to the Euclidean inner product on V^j. In the light of (4.51) the method is a multilevel relaxation in the direction of the truncated coarse grid functions

$$\lambda^k_{\mathcal{I},m} = P_{V^j_{\mathcal{I}}} \lambda^k_m = \sum_{i=1}^{n_j} \left(I_{\mathcal{I}} T_{j-1} \dots T_k\right)_{im} \lambda^j_i. \tag{4.53}$$

Note that these functions only appear implicitly and that their explicit construction is not necessary.

Remark 4.5. *The extension of the above multigrid method to V-cycles and W-cycles with multiple pre- and post-smoothing is straightforward. If one step of such linear truncated multigrid method is applied for the inexact solution of each linear subproblem in the nonsmooth Newton methods in Corollary 4.3 and 4.4 an overall nonlinear multigrid method is obtained. We call this method "Truncated Nonsmooth Newton Multigrid" and abbreviate this with TNNMG .*

The truncated basis functions have been introduced by Hoppe and Kornhuber [66] and further analyzed by Kornhuber and Yserentant [77]. In [69] they are used to accelerate the convergence of the monotone multigrid method for quadratic obstacle problems. The relation of the TNNMG method to this method will be considered in Section 4.3.2.

4.3.2 Relation to Multilevel Relaxation and Monotone Multigrid

We will now discuss the relation of the nonlinear TNNMG multigrid method introduced in the previous subsection to other nonlinear multigrid methods. First we concentrate on the quadratic obstacle problems in Example 4.1. For a detailed comparison of the presented algorithms for obstacle problems we refer to Gräser and Kornhuber [58].

Several multilevel methods to solve quadratic obstacle problems are based on the successive minimization of the energy in the directions given by the so-called multilevel nodal basis

$$\bigcup_{k=0}^{j} \Big\{ \lambda^k_i \in V^k : i = 1, \dots, n_k \Big\}.$$

As in the previous subsection the successive multilevel minimization algorithms can be expressed by grouping the directions with respect to their level k if we additionally introduce coarse representations $\underline{\psi}^k, \overline{\psi}^k$ of the obstacles:

```
$A_j = A,$
$r^j = b - A_j u^\nu,$
for $k = j, \dots, 0$ :
    for $i = 0, \dots, n_k$ :
        $\tilde{v}_i^k = (A_k)_{ii}^{-1} \left( r_i^k - \sum_{m=1}^{i-1} (A_k)_{im} v_m^k \right),$
        $v_i^k = \min\Big\{ \max\{ \tilde{v}_i^k, \underline{\psi}_i^k \}, \overline{\psi}_i^k \Big\},$
    if $k > 0$ :
        $r^{k-1} = T_{k-1}^T \left( r^k - A_k v^k \right),$
        $A_{k-1} = T_{k-1}^T A_k T_{k-1},$
$u^{\nu+1} = u^\nu + \sum_{k=0}^{j} (T_{j-1} \dots T_k) v^k.$
```

For a linear problem without constraints we have $v_i^k = \tilde{v}_i^k$ and this minimization procedure coincides with the multigrid algorithm presented in the previous subsection using the Gauß–Seidel smoother $B_k = D_k + L_k$. In this case the hierarchic formulation allows for an efficient implementation since all quantities needed to compute the correction v^k are available on the k-th level without considering higher levels besides the initial restriction of r^k and A_k from the next higher level. Depending on the selection of the coarse obstacles the latter need no longer be true for the constrained case.

The multilevel relaxation introduced by Mandel [80] uses exact coarse obstacles in the sense that the minimization of the functional J is done in the whole constraint set $K = \operatorname{dom} J$ for each coarse direction. This can be achieved using the coarse obstacles defined by

$$\begin{aligned} \underline{\psi}_i^k &= \min\Big\{ z \in \mathbb{R} : w^{k,i} + z \, (T_{j-1} \dots T_k) \, e_i^k \geq \underline{\psi} \Big\} \\ &= \max\Big\{ (\underline{\psi} - w^{k,i})_m / \left(T_{j-1} \dots T_k\right)_{mi} : m = 1, \dots, n_j \Big\}, \end{aligned}$$

$$\begin{aligned} \overline{\psi}_i^k &= \max\Big\{ z \in \mathbb{R} : w^{k,i} + z \, (T_{j-1} \dots T_k) \, e_i^k \leq \overline{\psi} \Big\} \\ &= \min\Big\{ (\overline{\psi} - w^{k,i})_m / \left(T_{j-1} \dots T_k\right)_{mi} : m = 1, \dots, n_j \Big\} \end{aligned}$$

where e_i^k is the i-th Euclidean basis vector in $\mathbb{R}^{n_k}$, $(T_{j-1} \dots T_k) \, e_i^k$ its level-j representation in $\mathbb{R}^{n_j}$, and $w^{k,i}$ is the level-j representation of the intermediate iterate before the i-th correction on level k, i.e.,

$$w^{k,i} = u^\nu + \sum_{m=k+1}^{j} (T_{j-1} \dots T_m)\, v^m + (T_{j-1} \dots T_k) \sum_{m=1}^{i-1} v_m^k e_m^k.$$

In order to compute these obstacles it is necessary to go up to the fine level for the computation of each scalar correction on all levels. For finite element discretizations this amounts in the evaluation of the defect obstacles

$$\underline{\psi} - w^{k,i}, \qquad \qquad \overline{\psi} - w^{k,i}$$

on the whole support of λ_i^k for each i, k. Thus the exact obstacles can only be used at the price of a higher complexity. Even for a problem resulting from uniform refinement this leads to $O(n_j)$ complexity for the k-th level. While the complexity is suboptimal the algorithm has been analyzed extensively. Based on a general result by Badea et al. [6] a polylogarithmic (with respect to the number of levels j) upper bound for the convergence rate was proved by Badea [5] for essentially uniformly refined grids.

One approach to avoid the complexity issue is to construct coarse obstacles $\underline{\psi}^k, \overline{\psi}^k \in \mathbb{R}^{n_k}$ such that the coarse constraint sets

$$K^k = \left\{ v \in \mathbb{R}^{n_k} : \underline{\psi}^k \leq v \leq \overline{\psi}^k \right\}$$

are subsets of the defect convex set, i.e.,

$$(T_{j-1} \dots T_k)\, K^k \subset \Big(K - u^\nu \cap (T_{j-1} \dots T_k)\, \mathbb{R}^{n_k} \Big),$$

and build an a priori decomposition of the defect convex set, i.e.,

$$K - u^\nu = \sum_{k=0}^{j} (T_{j-1} \dots T_k)\, K^k.$$

Tai [104] derived suitable hierarchic decompositions of $K - u^\nu$ that allow for similar convergence results as for the multilevel relaxation, while the algorithm has only $O(n_k)$ complexity on the k-th level. Since the a priori decomposition is very pessimistic, numerical examples show that this algorithm is by far slower then the multilevel relaxation using the optimal obstacles [58].

The monotone multigrid method introduced by Kornhuber [69] tries to overcome the complexity issue by using monotone restrictions of the defect obstacle that are neither recomputed for each scalar minimization problem nor completely a priori. Assuming that $(T_k)_{li} \geq 0$ and $\sum_i (T_k)_{li} = 1$ hold true (as it is the case for geometric multigrid) these restrictions are defined by

$$\begin{aligned} \underline{\psi}^j &= \underline{\psi} - u^\nu, & \underline{\psi}_i^k &= \max\Big\{ (\underline{\psi}^{k+1} - v^{k+1})_m : (T_k)_{mi} \neq 0 \Big\}, & k &= 0, \dots, j-1, \\ \overline{\psi}^j &= \overline{\psi} - u^\nu, & \overline{\psi}_i^k &= \min\Big\{ (\overline{\psi}^{k+1} - v^{k+1})_m : (T_k)_{mi} \neq 0 \Big\}, & k &= 0, \dots, j-1. \end{aligned}$$

The induced coarse constraint sets

$$K^k = \left\{ v \in \mathbb{R}^{n_k} : \underline{\psi}^k \leq v \leq \overline{\psi}^k \right\}$$

are subsets of the defect convex set incorporating all fine level corrections, i.e.,

$$(T_{j-1} \dots T_k)\, K^k \subset \Big(K - u^\nu - \sum_{m=k+1}^{j} (T_{j-1} \dots T_m)\, v^m \cap (T_{j-1} \dots T_k)\, \mathbb{R}^{n_k} \Big),$$

While this simplification allows to retain $O(n_k)$ complexity on the k-th level, only asymptotic bounds on the convergence rate are known [69]. In practice this algorithm shows almost the same convergence rates as the multilevel relaxation with optimal obstacles [58].

For non-degenerate problems the above algorithms satisfy

$$\{i : u_i^\nu = \underline{\psi}_i\} = \{i : u_i^* = \underline{\psi}_i\}, \qquad \{i : u_i^\nu = \overline{\psi}_i\} = \{i : u_i^* = \overline{\psi}_i\}, \qquad \forall \nu > \nu_0$$

for some finite iteration step ν_0. From this moment on the multilevel relaxation as well as the monotone multigrid method reduce to linear multigrid methods for a linear problem restricted to all inactive components. However, their convergence will in general be much slower than for a similar problem without constraints. The main reason is that all coarse directions containing active components cannot contribute to corrections since this would violate the constraints.

In order to accelerate the method the use of truncated coarse grid functions was suggested by Hoppe and Kornhuber [66], Kornhuber [69]. For the truncated monotone multigrid method the current inactive set after the application of the fine grid smoother is defined by

$$\mathcal{I}_\nu = \{i : u^\nu + v^j\}.$$

The only difference to the untruncated version is that the transfer operator T_{j-1} is replaced by the truncated version $\mathcal{I}_\nu T_{j-1}$ and that $\tilde{v}_i^k$ is set to zero if $(A_k)_{ii}$ is zero. As noted in the previous subsection this is equivalent to replacing the standard coarse grid functions by their truncated versions defined in (4.53). Similar to the standard monotone multigrid method this algorithm degenerates to a linear multigrid method with truncated basis functions. The convergence rates of such linear methods were analyzed by Kornhuber and Yserentant [77] and carry over to asymptotic rates for the truncated monotone multigrid method [72]. By the same arguments as in Section 4.2.5 one can, however, not expect global mesh-independent convergence.

Now we compare the truncated monotone multigrid method (TMMG) with the TNNMG method with nonlinear Gauß–Seidel smoother. Remember that the latter is obtained if a linear multigrid step with truncated basis functions is applied to the linear subproblems in the algorithm in Corollary 4.3. First we note that the nonlinear smoothing step and the selected inactive set in the TNNMG method coincides with the nonlinear fine level smoothing and the inactive set in the TMMG method, respectively, i.e.,

$$u^{\nu+\frac{1}{2}} = u^\nu + \mathcal{F}_{GS}(u^\nu) = u^\nu + v^j, \qquad\qquad \mathcal{I}''(u^{\nu+\frac{1}{2}}) = \mathcal{I}_\nu.$$

Furthermore, the TMMG method ensured feasibility and monotonicity of all intermediate iterates by using coarse grid obstacles while the TNNMG method allows the violation for intermediate iterates and ensures feasibility and monotonicity by a projection and subsequent line search of the whole coarse correction. The only further difference is that the TNNMG method does also apply a linear fine grid smoother.

This close relation can also be seen numerically. For many problems both methods behave almost the same [58] despite the fact that the line search in the TNNMG method often leads to an overrelaxation that accelerates convergence. In special cases, where the monotone projection leads to very restrictive coarse obstacles (e.g. contact problems in elasticity with complicated domains), the TNNMG method clearly outperforms the TMMG method (see [62]). Another advantage of the TNNMG method is that it does only need a linear multigrid method. Thus the extension to W-cycles and the usage of other smoothers is straight forward.

Remark 4.6. *For a quadratic energy J_0 and other choices of φ the TNNMG method is closely related to the constraint Newton linearization method introduced by Kornhuber [75]. This method also relies on a nonlinear Gauß–Seidel smoother on the finest level and a Newton-type coarse grid correction. The coarse correction is based on essentially the same truncated linearization.*

In contrast to the projection used in the TNNMG method, feasibility of the coarse correction is guaranteed by coarse obstacles that restrict the correction to a region where φ is smooth. The obtained quadratic obstacle problems are solved approximately using one step of the truncated monotone multigrid method.

Monotonicity of the coarse correction is ensured by a priori computed local damping parameters for each coarse grid basis function instead of a global line search used by the TNNMG method for the whole correction.

5 Schur Nonsmooth Newton Methods for Set-Valued Saddle Point Problems

While the problems considered in the previous chapter only incorporate local componentwise constraints we now consider additional global linear constraints. Such problems can be written as saddle point problems with Lagrangian multipliers for the linear constraints. Since the energy is still allowed to contain nonsmooth nonlinearities these saddle point problems will in general be inclusions for set-valued operators. This chapter is dedicated to the development of an efficient algebraic solver for this problem class.

We start by introducing the saddle point problem and derive a dual minimization problem whose optimality system is a nonlinear Schur complement equation. Then we present and analyze descent algorithms for the dual problem including inexact versions. In order to accelerate these methods we derive generalized linearizations for the nonlinear Schur complement that can be used in a Newton-type versions of the descent algorithms. Finally we discuss the convergence properties and some computational aspects of the obtained Schur Nonsmooth Newton method and its relation to other algorithms for certain special cases of the considered problem class.

5.1 Nonsmooth Convex Minimization Problems with Linear Constraints

Throughout this chapter we consider the nonlinear saddle point problem

$$u^* \in \mathbb{R}^n,\ w^* \in \mathbb{R}^m : \qquad \begin{pmatrix} F & B^T \\ B & -C \end{pmatrix} \begin{pmatrix} u^* \\ w^* \end{pmatrix} \ni \begin{pmatrix} f \\ g \end{pmatrix} , \tag{5.1}$$

where B, C are suitable matrices, and the set-valued operator $F = \partial J$ is the subdifferential of a strictly convex functional $J : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. If not stated otherwise J is assumed to be of the form $J = J_0 + \varphi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ as introduced in Chapter 4 with J_0 and φ satisfying (A5) and (A8), respectively. Furthermore, we assume:

(A10) $B \in \mathbb{R}^{m,n}$, $f \in \mathbb{R}^n$, and $g \in \mathbb{R}^m$. $C \in \mathbb{R}^{m,m}$ is symmetric and positive semidefinite.

(A11) The saddle point problem (5.1) has a unique solution.

For $C = 0$ this problem is equivalent to the minimization of $J - \langle f, \cdot \rangle$ subject to the linear constraint $Bu = g$. We are interested in the fast solution of this class of problems. Classical Newton linearization (see, e.g., Deuflhard [43], [83, 102]) can in general not be used because of possible nonsmoothness of the nonlinearity φ. For example, Problem 3.18 and Problem 3.19 obtained in Chapter 3 for discretized Cahn–Hilliard equations contain the logarithmic potential that degenerates rapidly to an indicator functional if the temperature goes to zero. Although this potential is differentiable inside of $(-1, 1)$ its limiting properties make it de facto nonsmooth for small temperatures. On the other hand primal [65] and primal–dual [14, 64] active set methods are restricted to special cases where the nonlinearity is related to inequality constraints. Furthermore, they do not use the inherent convex structure to achieve global convergence. Monotone multigrid methods for obstacle problems [69], their extension by constrained Newton linearization to nonsmooth nonlinear problems [71, 74, 75], and the TNNMG method derived in the previous chapter are known to be efficient globally convergent methods. However, they cannot deal with the linear constraints in the given saddle point problems.

Before we develop a new method for the efficient iterative solution of this problem class we derive an equivalent dual minimization problem.

Proposition 5.1. *Let $J : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a strictly convex, proper and lower semicontinuous function whose subdifferential $F = \partial J$ has a single-valued Lipschitz continuous inverse F^{-1}. Furthermore, assume that (A10) and (A11) are true. Then the saddle point problem* (5.1) *is equivalent to*

$$w^* \in \mathbb{R}^m : \qquad H(w^*) = 0 \tag{5.2}$$

with the Lipschitz continuous, monotone operator $H : \mathbb{R}^m \to \mathbb{R}^m$ given by

$$H(w) = -BF^{-1}(f - B^T w) + Cw + g \,, \qquad w \in \mathbb{R}^m \,. \tag{5.3}$$

Proof. Due to the properties of J and F straightforward block elimination in (5.1) provides the equivalence. Since H consists of a sum and a composition of F^{-1} with affine functions the Lipschitz continuity follows directly from the Lipschitz continuity of F^{-1}.

The convexity of J implies the monotonicity of F^{-1}. In combination with the nonnegativity of C we get

$$\begin{aligned}
&\langle w_1 - w_2, H(w_1) - H(w_2)\rangle \\
&\quad = \big\langle (f - B^T w_1) - (f - B^T w_2), F^{-1}(f - B^T w_1) - F^{-1}(f - B^T w_2)\big\rangle \\
&\qquad\qquad + \langle C(w_1 - w_2), w_1 - w_2\rangle \geq 0,
\end{aligned} \tag{5.4}$$

yielding monotonicity of H. □

The operator H can be regarded as a nonlinear Schur complement. For a linear saddle point problem with $F = A$ for some symmetric positive definite matrix A it reduces to $H(w) = (BA^{-1}B^T + C)w + (g - BA^{-1}f)$. The part $BA^{-1}B^T$ of this operator is the usual linear Schur complement. In contrast to the linear case, the right hand side f cannot be separated from the part depending on w in general. Note that although the saddle point problem is set-valued, the operator H is single-valued, because $F^{-1} = (\partial\varphi)^{-1}$ is single-valued or, equivalently, the minimization of $J(x) - \langle y, x\rangle$ on $\mathbb{R}^n$ admits a unique solution.

Theorem 5.1. *Under the assumptions of Proposition 5.1 there is a Fréchet-differentiable, convex functional $h : \mathbb{R}^m \to \mathbb{R}$ with the property $\nabla h = H$ and the representation*

$$h(w) = -\mathcal{L}(F^{-1}(f - B^T w), w)\ , \qquad w \in \mathbb{R}^m\ , \tag{5.5}$$

where

$$\mathcal{L}(u, w) = J(u) - \langle f, u\rangle + \langle Bu - g, w\rangle - \frac{1}{2}\langle Cw, w\rangle$$

denotes the Lagrange functional associated with the saddle point problem (5.1)*.*

Proof. By [49, Corollary 5.2, p. 22] the polar (or conjugate) functional

$$J^*(y) = \sup_{x\in\mathbb{R}^n} (\langle y, x\rangle - J(x)) = -\inf_{x\in\mathbb{R}^n} (J(x) - \langle y, x\rangle)$$

of J is convex and has the property $\partial J^* = (\partial J)^{-1} = F^{-1}$. Since $F^{-1}y$ is single-valued for all $y \in \mathbb{R}^n$ its polar J^* can take only finite values and the domain of the polar is $\mathbb{R}^n$. Thus J^* is continuous on the whole space $\mathbb{R}^n$ by [49, Corollary 2.3, p. 12].

By [49, Proposition 5.3, p. 23] finiteness and continuity of J^*, and single-valuedness of ∂J^* imply Gâteaux-differentiability of J^*. The continuity of $\partial J^* = F^{-1}$ implies that J^* is even Fréchet-differentiable with $\nabla J^* = F^{-1}$. Setting

$$h(w) = J^*(f - B^T w) + \frac{1}{2}\langle Cw, w\rangle + \langle g, w\rangle \tag{5.6}$$

we immediately get $\nabla h = H$ using the chain rule. Convexity of h directly follows from convexity of J^*, and symmetry and positivity of C. Finally, inserting

$$J^*(y) = -\big(J(F^{-1}(y)) - \langle y, F^{-1}(y)\rangle\big)$$

with $y = f - B^T w$ into (5.6) gives (5.5). □

As immediate consequence of Proposition 5.1 and Theorem 5.1 we get the equivalence of (5.1) to an unconstrained dual problem.

Corollary 5.1. *The set-valued saddle point problem* (5.1) *is equivalent to the dual unconstrained convex minimization problem*

$$w^* \in \mathbb{R}^m : \qquad h(w^*) \le h(w) \qquad \forall w \in \mathbb{R}^m. \tag{5.7}$$

By the arguments in the proof of Theorem 5.1 not only h but also $\tilde{h}(w) = J^*(f - B^T w) + \langle g, w\rangle$ is convex. Hence

$$h(\lambda x + (1-\lambda)y) \le \lambda h(x) + (1-\lambda)h(y) - \lambda(1-\lambda)\frac{1}{2}\|x - y\|_C^2 \qquad \forall \lambda \in [0,1] \tag{5.8}$$

holds for all $x, y \in \mathbb{R}^m$, which means that h is strongly convex if C is symmetric and positive definite. However, the latter need not be the case, and h is in general not even strictly convex so that we have to require uniqueness separately.

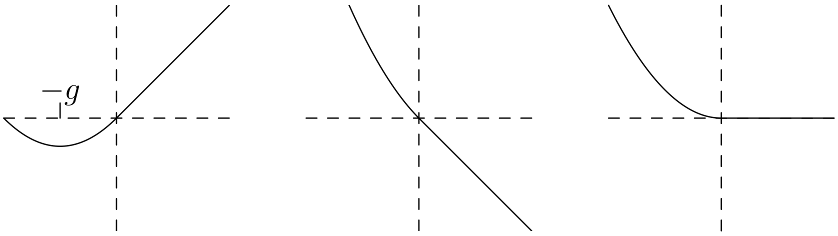


Figure 5.1: The dual functional h in Example 5.1: Unique minimizer at $-g$ for $g > 0$ (left), no minimizer for $g < 0$ (middle), non unique minimizer for $g = 0$ (right).

Example 5.1. *Consider the saddle point problem*

$$u, w \in \mathbb{R}: \qquad \begin{pmatrix} 1 + \partial\chi_{[0,\infty)} & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} u \\ w \end{pmatrix} \ni \begin{pmatrix} 0 \\ g \end{pmatrix} \tag{5.9}$$

for some $g \in \mathbb{R}$. Then the dual functional is given by

$$h(w) = \begin{cases} \frac{1}{2}w^2 + gw & \text{if } w \leq 0, \\ gw & \text{if } w > 0, \end{cases}$$

and is not strictly convex on $(0, \infty)$. We have to distinguish three cases (see Figure 5.1):

- *For $g > 0$ the solution of (5.9) is uniquely given by $(u, w) = (g, -g)$ and thus the minimizer $w = -g$ of h is also unique.*
- *For $g < 0$ the linear constraint does still imply $u = g$ but $\partial\chi_{[0,\infty)}(g)$ is empty. Hence there is no solution of (5.9). Conversely h has no minimizer since it is not bounded from below in this case.*
- *For $g = 0$ all pairs (u, w) with $u = 0$ and $w \geq 0$ are solutions to (5.9) and all such w minimize h.*

Corollary 5.1 offers the possibility to treat the nonsmooth saddle point problem (5.1) as a smooth unconstrained minimization problem or as an operator equation with a Lipschitz continuous monotone operator. This simplification comes at the price of the fact that the functional h and the operator $H = \nabla h$ might be expensive to evaluate, since both involve the evaluation of $F^{-1} = (\partial J)^{-1}$ and thus the solution of an unconstrained minimization problem for the nonsmooth functional J.

5.2 Descent Methods for the Dual Problem

Once we have reformulated the saddle point problem (5.1) as the dual minimization problem (5.7), descent methods for unconstrained minimization of differentiable functionals can be applied.

Since the operator F^{-1} involved in h and $H = \nabla h$ is in general not directly available, it is complicated and expensive to use iterative methods based on local properties or substeps as e.g. the Gauß–Seidel or Jacobi method for the solution of (5.7). For this reason we consider so-called gradient-related algorithms based on global descent directions. Although there are numerous basic convergence proofs (see, e.g., the classic text book by Ortega and Rheinboldt [83]) for this class of methods none of them can be applied directly to our setting. Thus we extend the results in a way fitting in the presented framework.

Throughout this section we assume that $h : \mathbb{R}^m \to \mathbb{R}$ is a continuously differentiable convex function having a unique minimizer w^* and a Lipschitz continuous derivative ∇h.

The results are presented in terms of the norm $\|\cdot\|_M$,

$$\|x\|_M^2 = \langle Mx, x\rangle \ , \qquad x \in \mathbb{R}^m \ ,$$

induced by a symmetric positive definite matrix $M \in \mathbb{R}^{m,m}$. Elements x' of the dual space $(\mathbb{R}^m)'$ will be represented as $x' = \langle x, \cdot\rangle$ with suitable $x \in \mathbb{R}^m$. In view of

$$|x'(y)| = |\langle x, y\rangle| \leq \|M^{-\frac{1}{2}}x\| \|M^{\frac{1}{2}}y\| = \|x\|_{M^{-1}} \|y\|_M \ ,$$

the dual space $(\mathbb{R}^m, \|\cdot\|_M)'$ is identified with $(\mathbb{R}^m, \|\cdot\|_{M^{-1}})$.

Gradient related descent methods are of the form

$$w^{\nu+1} = w^\nu + \rho_\nu d^\nu, \qquad \nu = 1, \ldots \tag{5.10}$$

for a given initial iterative w^0. In each step, first a search direction d^ν is chosen according to the current iterate w^ν. Then, a step size ρ_ν is fixed depending on w^ν and d^ν, i.e.,

$$d^\nu = d(\nu, w^\nu), \qquad \rho_\nu = \rho(\nu, w^\nu, d^\nu) \ , \quad \nu = 0, 1, \ldots \tag{5.11}$$

with suitable mappings d and ρ. It will turn out that monotonicity is again the crucial property. Having this in mind and in view of the algorithms developed in Chapter 4 we consider the extended algorithm

$$w^{\nu+\frac{1}{2}} = w^\nu + \rho_\nu d^\nu, \tag{5.12}$$

$$w^{\nu+1} = w^{\nu+\frac{1}{2}} + \mathcal{C}(w^{\nu+\frac{1}{2}}) \tag{5.13}$$

with an operator $\mathcal{C}$ having the property $h(w + \mathcal{C}(w)) \leq h(w)$.

5.2.1 Convergence Analysis

In order to obtain a convergent method the descent directions should allow for sufficient descent of h and the step sizes must realize the descent.

Definition 5.1. *The map $d : \mathbb{N} \times \mathbb{R}^m \to \mathbb{R}^m$ is said to generate descent directions if for any sequence $w^\nu \subset \mathbb{R}^m$ the directions $d^\nu = d(\nu, w^\nu)$ satisfy*

$$\nabla h(w^\nu) = 0 \iff d^\nu = 0, \qquad \forall \nu \in \mathbb{N} \tag{5.14}$$

and

$$\nabla h(w^\nu) \neq 0 \quad \Rightarrow \quad \langle \nabla h(w^\nu), d^\nu \rangle < 0, \qquad \forall \nu \in \mathbb{N}. \tag{5.15}$$

If $d : \mathbb{N} \times \mathbb{R}^m \to \mathbb{R}^m$ generates descent directions we will also call the generated sequence d^ν a sequence of descent directions. Although a sequence of descent directions allows for descent in each iteration step the method might not converge if the directions degenerate in the sense that the angles between d^ν and $\nabla h(w^\nu)$ tend to $\frac{\pi}{2}$. Thus we have to impose a stronger condition than (5.15) to exclude this case.

Definition 5.2. *The map $d : \mathbb{N} \times \mathbb{R}^m \to \mathbb{R}^m$ is said to generate gradient-related directions, if for any sequence $w^\nu \subset \mathbb{R}^m$ the directions $d^\nu = d(\nu, w^\nu)$ satisfy*

$$\nabla h(w^\nu) = 0 \iff d^\nu = 0, \qquad \forall \nu \in \mathbb{N} \tag{5.16}$$

and

$$-\langle \nabla h(w^\nu), d^\nu \rangle \geq c_D \left\| \nabla h(w^\nu) \right\|_{M^{-1}} \left\| d^\nu \right\|_M, \qquad \forall \nu \in \mathbb{N} \tag{5.17}$$

with a constant $c_D > 0$ independent of ν.

Note that the preconditioned gradients $d(\nu, w^\nu) = -M^{-1} \nabla h(w^\nu)$ are gradient-related since (5.17) is satisfied with equality and $c_D = 1$.

Definition 5.3. *Let $d : \mathbb{N} \times \mathbb{R}^m \to \mathbb{R}^m$ generate descent directions. Then $\rho : \mathbb{N} \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ is said to generate efficient step sizes, if for any sequence $w^\nu \subset \mathbb{R}^m$ and $d^\nu = d(\nu, w^\nu)$ the step sizes $\rho_\nu = \rho(\nu, w^\nu, d^\nu)$ satisfy*

$$h(w^\nu + \rho_\nu d^\nu) \leq h(w^\nu) - c_S \left(\frac{\langle \nabla h(w^\nu), d^\nu \rangle}{\left\| d^\nu \right\|_M} \right)^2 \tag{5.18}$$

for all $\nu \in \mathbb{N}$ such that $\nabla h(w^\nu) \neq 0$ with a constant $c_S > 0$ independent of ν.

In order to prove convergence it will be necessary to exploit a compactness property provided by the following lemma.

Lemma 5.1. *The sublevel set $\{w \in \mathbb{R}^m : h(w) \leq C\}$ is compact for every $C \geq h(w^*)$.*

Proof. Continuity of h implies that the set is closed. Assume that $S_{h \leq c} := \{w \in \mathbb{R}^m : h(w) \leq c\}$ is not bounded for some $c \geq h(w^*)$. Then there exists a sequence $(w^k) \subset S_{h \leq h(w^0)}$ with the property $\|w^k - w^*\| \geq k$. We define the bounded sequence

$$\tilde{w}^k = w^* + (w^k - w^*)/\|w^k - w^*\|$$

contained in the unit sphere with center w^*. Compactness guarantees the existence of a convergent subsequence $\tilde{w}^{k_j}$. By continuity and convexity of h it satisfies

$$\lim_{j\to\infty} h(\tilde{w}^{k_j}) \le \lim_{j\to\infty} k_j^{-1} h(w^{k_j}) + (1 - k_j^{-1}) h(w^*) = h(w^*).$$

Finally uniqueness of w^* implies $\tilde{w}^{k_j} \to w^*$ contradicting $\|\tilde{w}^{k_j} - w^*\| = 1$. □

Lemma 5.2. *Let $w^\nu \in \mathbb{R}^m$ be a sequence such that $\nabla h(w^\nu) \to 0$ and $h(w^\nu) \le C$ hold for some $C \ge h(w^*)$. Then w^ν converges to w^*.*

Proof. By Lemma 5.1 the set $S = \{w \in \mathbb{R}^m : h(w) \le C\}$ is compact. As a consequence, there is a convergent subsequence of w^ν.

Now let $w^{\nu_i} \to w^{**}$ be any convergent subsequence with the limit $w^{**} \in \mathbb{R}^m$. Then the continuity of ∇h provides $\nabla h(w^\nu) \to \nabla h(w^{**}) = 0$. Uniqueness of the minimizer proves the assertion. □

We are now ready to prove that the combination of gradient-related descent directions and efficient step sizes leads to a globally convergent method. Although this is a standard result that (with small modifications) can be found in many textbooks (see, e.g., [53, 83, 102]), we give a proof here since these variant do for example not include the monotone correction $\mathcal{C}$.

Theorem 5.2. *Assume that d and ρ generate gradient-related directions and efficient step sizes, respectively. Then the iterates w^ν generated by* (5.11), (5.12), *and* (5.13) *converge to w^* for an arbitrary initial iterate $w^0 \in \mathbb{R}^m$.*

Proof. Combining the properties of $d^\nu = d(\nu, w^\nu)$, $\rho_\nu = \rho(\nu, w^\nu, d^\nu)$, and $\mathcal{C}$ we get

$$h(w^\nu) - h(w^{\nu+1}) \ge h(w^\nu) - h(w^{\nu+\frac{1}{2}}) \ge c_S c_D^2 \|\nabla h(w^\nu)\|_{M^{-1}}^2 \qquad \forall \nu \in \mathbb{N}. \tag{5.19}$$

Since h has a global minimizer, the sequence $h(w^\nu)$ is bounded from below and, by (5.19), monotonically decreasing. Hence, $h(w^\nu)$ converges to some $h^* \in \mathbb{R}$ and (5.19) implies

$$\|\nabla h(w^\nu)\|_{M^{-1}}^2 \to 0. \tag{5.20}$$

Together with $h(w^\nu) \le h(w^0)$ this allows to apply Lemma 5.2 which proves the assertion. □

Under the stronger assumption that h is strongly convex, i.e., if there is a $\mu > 0$ such that

$$h(\lambda x + (1-\lambda)y) \le \lambda h(x) + (1-\lambda)h(y) - \lambda(1-\lambda)\frac{\mu}{2}\|x - y\|_M^2 \qquad \forall \lambda \in [0,1],$$

we get R-linear convergence. The following results of Lemma 5.3 and Theorem 5.3 can even be extended to infinite-dimensional problems without additional assumptions.

Lemma 5.3. *Let h be strongly convex with a constant $\mu > 0$. Then the following estimates holds*

$$\frac{\mu}{2} \|w - w^*\|_M^2 \leq h(w) - h(w^*) \leq \frac{1}{2\mu} \|\nabla h(w)\|_{M^{-1}}^2 \qquad \forall w \in \mathbb{R}^m. \tag{5.21}$$

Proof. By Lemma A.1 in the appendix strong convexity implies

$$h(x) - h(y) \geq \langle \nabla h(y), x - y \rangle + \frac{\mu}{2} \|x - y\|_M^2 \qquad \forall x, y \in \mathbb{R}^m.$$

Inserting $x = w$ and $y = w^*$ gives the left inequality while inserting $x = w^*$ and $y = w$ together with Young's inequality provides the right one. □

Theorem 5.3. *Assume that the conditions of Theorem 5.2 and Lemma 5.3 are satisfied. Then the iterates w^ν generated by* (5.11), (5.12), *and* (5.13) *satisfy the error estimate*

$$\|w^\nu - w^*\|_M^2 \leq q^\nu \frac{2}{\mu} \left(h(w^0) - h(w^*) \right) \tag{5.22}$$

with $q = (1 - 2c_S c_D^2 \mu) < 1$.

Proof. By (5.12), (5.13), and the right inequality in (5.21) we get

$$\begin{aligned} 0 \leq h(w^{\nu+1}) - h(w^*) &\leq h(w^\nu) - h(w^*) + h(w^{\nu+\frac{1}{2}}) - h(w^\nu) \\ &\leq h(w^\nu) - h(w^*) - c_S c_D^2 \|\nabla h(w^\nu)\|_{M^{-1}}^2 \\ &\leq (1 - c_S c_D^2 2\mu)(h(w^\nu) - h(w^*)). \end{aligned}$$

Combining this estimate with the left inequality in (5.21) yields

$$\frac{\mu}{2} \|w^\nu - w^*\|_M^2 \leq h(w^\nu) - h(w^*) \leq (1 - c_S c_D^2 2\mu)^\nu (h(w^0) - h(w^*)),$$

which proves the assertion. Note that for $w^0 \neq w^*$ the estimate implies $0 \leq 1 - c_S c_D^2 2\mu$ because $h(w^\nu) - h(w^*) > 0$. □

5.2.2 Inexact Evaluation of Descent Directions

We now consider inexact search directions $\tilde{d}^\nu$ obtained if the exact evaluation $d^\nu = d(\nu, w^\nu)$ is replaced by some approximation

$$\tilde{d}^\nu = \tilde{d}(\nu, w^\nu) \approx d(\nu, w^\nu). \tag{5.23}$$

Proposition 5.2. *Let d generate gradient-related directions that satisfy* (5.17) *with the constant c_D, and let $\tilde{d}$ generate descent directions. Assume that there is a constant $c < c_D/2$ such that the approximations $\tilde{d}^\nu = \tilde{d}(\nu, w^\nu)$ satisfy at least one of the accuracy conditions*

$$\|d^\nu - \tilde{d}^\nu\|_M \leq c \|\tilde{d}^\nu\|_M \quad \forall \nu \in \mathbb{N}, \tag{5.24}$$

$$\|d^\nu - \tilde{d}^\nu\|_M \leq c \|d^\nu\|_M \quad \forall \nu \in \mathbb{N}, \tag{5.25}$$

for all sequences w^ν. Then the approximation $\tilde{d}$ does also generate gradient-related directions that satisfy (5.17) *with the constant $\tilde{c}_D = c_D - 2c > 0$.*

Proof. Let $w^\nu \subset \mathbb{R}^m$. Then the vectors $d^\nu = d(\nu, w^\nu)$ satisfy (5.16) and (5.17). We have to prove a similar estimate for the approximations $\tilde{d}^\nu$. For $\tilde{d}^\nu = 0$ this is trivial. For $\tilde{d}^\nu \neq 0$ we instantly get $\nabla h(w^\nu) \neq 0$ and thus $d^\nu \neq 0$. In this case elementary calculations involving the Cauchy–Schwarz inequality and the triangle inequality yield

$$\left| \left\langle \frac{\nabla h(w^\nu)}{\|\nabla h(w^\nu)\|_{M^{-1}}}, \frac{d^\nu}{\|d^\nu\|_M} - \frac{\tilde{d}^\nu}{\|\tilde{d}^\nu\|_M} \right\rangle \right| \leq 2 \frac{\|d^\nu - \tilde{d}^\nu\|_M}{\|\tilde{d}^\nu\|_M} .$$

From $\|d^\nu - \tilde{d}^\nu\|_M / \|\tilde{d}^\nu\|_M \leq c < c_D/2$ we get

$$- \left\langle \nabla h(w^\nu), \tilde{d}^\nu \right\rangle \geq \tilde{c}_D \|\nabla h(w^\nu)\|_{M^{-1}} \|\tilde{d}^\nu\|_M$$

with $\tilde{c}_D = c_D - 2c > 0$. The proof for the second accuracy condition (5.25) uses exactly the same arguments with $\|d^\nu - \tilde{d}^\nu\|_M / \|\tilde{d}^\nu\|_M$ replaced by $\|d^\nu - \tilde{d}^\nu\|_M / \|d^\nu\|_M$. □

Since the constant c_D needed to check the accuracy conditions in Proposition 5.2 with $c < c_D/2$ is in general not known, we replace them by the asymptotic criteria

$$\lim_{\nu \to \infty} \frac{\|d^\nu - \tilde{d}^\nu\|_M}{\|\tilde{d}^\nu\|_M} = 0 \qquad \text{and} \qquad \lim_{\nu \to \infty} \frac{\|d^\nu - \tilde{d}^\nu\|_M}{\|d^\nu\|_M} = 0, \tag{5.26}$$

respectively. They imply that the criteria in Proposition 5.2 with $c < c_D/2$ hold for sufficiently large ν with arbitrarily small c. To see that the whole sequence $\tilde{d}^\nu$ is gradient-related assume that (5.24) or (5.25) is satisfied for $\nu > \nu_0$. Hence by Proposition 5.2 the estimate (5.17) holds for $\nu > \nu_0$ with $\tilde{c}_D$. Then it also holds for all ν with the constant

$$\tilde{\tilde{c}}_D = \min \left\{ \tilde{c}_D, \min \left\{ - \frac{\langle \nabla h(w^\nu), d^\nu \rangle}{\|\nabla h(w^\nu)\|_{M^{-1}} \|d^\nu\|_M} : \nu \leq \nu_0 \right\} \right\} > 0.$$

Furthermore, the constants $\tilde{c}_D, \tilde{\tilde{c}}_D$ in estimate (5.17) for $\tilde{d}^\nu$ tend to the constant c_D for the exact directions d^ν in this case.

Remark 5.1. *Another approach allows to even use almost arbitrarily inexact evaluation. Having the choice to add a monotone correction $\mathcal{C}$ in* (5.13) *offers the possibility to change the role of the descent direction according to*

$$\begin{aligned} w^{\nu+\frac{1}{2}} &= w^\nu - \rho_\nu \nabla h(w^\nu), \\ w^{\nu+1} &= w^{\nu+\frac{1}{2}} + \rho_{\nu+\frac{1}{2}} \tilde{d}^{\nu+\frac{1}{2}}. \end{aligned}$$

Since the directions $-\nabla h(w^\nu)$ are trivially gradient-related we only have to impose the condition that ρ_ν is efficient and that $\rho_{\nu+1/2}$ is such that $h(w^{\nu+1}) \leq h(w^{\nu+\frac{1}{2}})$. Then Theorems 5.2 and 5.3 can be applied without any accuracy condition on $\tilde{d}^{\nu+\frac{1}{2}}$. Convergence is guaranteed by the leading gradient step.

A possible drawback of this approach is the additional computational effort. It amounts to the evaluation of $-\nabla h(w^\nu)$ and the computation of the second damping parameter $\rho_{\nu+1/2}$. The later can be less expensive then the computation of ρ_ν since it only has to provide monotonicity. However, if the evaluation of F^{-1} is expensive it might still be too costly.

5.2.3 Step Size Rules

There is a multitude of algorithms for the selection of efficient step sizes available from textbooks and surveys like [43, 82, 83, 87]. Since it is probably the most common we present the step size rule by Armijo [4] (see also [43, 83]) which tracks the actual decrease of the functional h. Furthermore, we show that the inexact evaluation of the so-called "exact step sizes", e.g. by bisection, also leads to efficient step sizes. This approach relies on the derivatives of h instead of its values. Finally we prove convergence for the sequence obtained if the step rule is adaptively switched on and off in each step using a simple criterion depending on d^ν only.

Both step size rules consider the function h only along the line spanned by the current iterate w^ν and the descent direction d^ν. Thus we define the scalar, convex function $\psi : [0, \infty) \to \mathbb{R}$ by

$$\psi(r) = h(w^\nu + rd^\nu).$$

Note that its derivative

$$\psi'(r) = \langle \nabla h(w^\nu + rd^\nu), d^\nu \rangle$$

is Lipschitz continuous with Lipschitz constant $L \|d^\nu\|_M^2$ where L denotes the Lipschitz constant of ∇h, i.e.,

$$\|\nabla h(v) - \nabla h(w)\|_{M^{-1}} \leq L \|v - w\|_M \qquad \forall v, w \in \mathbb{R}^m \, . \tag{5.27}$$

In order to define the Armijo step sizes consider a fixed parameter $\delta \in (0, 1)$. Then a step size $r \geq 0$ is called "admissible" if

$$h(w^\nu + rd^\nu) \leq h(w^\nu) + r\delta \langle \nabla h(w^\nu), d^\nu \rangle \tag{5.28}$$

is satisfied.

Proposition 5.3. *Let $d : \mathbb{N} \times \mathbb{R}^m \to \mathbb{R}^m$ generate descent directions. For a sequence $w^\nu \subset \mathbb{R}^m$ and directions $d^\nu = d(\nu, w^\nu)$ assume that fixed parameters $\alpha > 0$ and $\delta, \beta \in (0, 1)$ and a sequence α_ν with*

$$\alpha_\nu \geq -\alpha \frac{\langle \nabla h(w^\nu), d^\nu \rangle}{\|d^\nu\|_M^2}$$

is given. Then the Armijo rule defined by

$$\rho(\nu, w^\nu, d^\nu) := \max \left\{ r = \alpha_\nu \beta^j \ : r \text{ admissible } , j \in \mathbb{N} \cup \{0\} \right\} \qquad \forall d^\nu \neq 0 \tag{5.29}$$

and $\rho(\nu, w^\nu, 0) := 0$ generates efficient step sizes. More precisely the step sizes $\rho_\nu = \rho(\nu, w^\nu, d^\nu)$ satisfy (5.18) with

$$c_S = \delta \min \left\{ \alpha, \beta \left(\frac{1 - \delta}{L} \right) \right\} \, . \tag{5.30}$$

Proof. Let $\nu \in \mathbb{N}$ such that $d^\nu \neq 0$. Then $\psi'(0) = \langle \nabla h(w^\nu), d^\nu \rangle < 0$ holds by (5.14) and (5.15). The admissibility condition (5.28) can be rewritten as

$$\psi(r) \leq \psi(0) + r\delta\psi'(0) \; . \tag{5.31}$$

The step size ρ_ν is well-defined, because the set appearing in (5.29) is not empty. Otherwise, we would have

$$\psi(\alpha_\nu \beta^j) > \psi(0) + \delta\alpha_\nu\beta^j\psi'(0) \qquad \forall j \in \mathbb{N} \; ,$$

providing

$$\psi'(0) = \lim_{j\to\infty} \frac{\psi(\alpha_\nu\beta^j) - \psi(0)}{\alpha_\nu\beta^j} \geq \delta\psi'(0)$$

and, as $\psi'(0) < 0$, the contradiction $\delta \geq 1$.

For the proof of (5.18) we have to distinguish two cases. If $\rho_\nu = \alpha_\nu$, i.e., $j = 0$, then the estimate

$$\psi(\rho_\nu) \leq \psi(0) - \delta\alpha \left(\frac{\psi'(0)}{\|d^\nu\|_M} \right)^2$$

holds trivially by (5.29). Now let $\rho_\nu = \alpha_\nu\beta^j$ with $j > 0$. Then

$$\psi\left(\frac{\rho_\nu}{\beta}\right) > \psi(0) + \delta\frac{\rho_\nu}{\beta}\psi'(0)$$

and thus for some $\xi \in [0,1]$

$$\frac{\rho_\nu}{\beta}\psi'\left(\xi\frac{\rho_\nu}{\beta}\right) = \psi\left(\frac{\rho_\nu}{\beta}\right) - \psi(0) > \delta\frac{\rho_\nu}{\beta}\psi'(0).$$

In combination with the Lipschitz continuity (5.27) this leads to

$$-(1-\delta)\psi'(0) \;\leq\; \psi'\left(\xi\frac{\rho_\nu}{\beta}\right) - \psi'(0) \;\leq\; L\xi\frac{\rho_\nu}{\beta}\|d^\nu\|_M^2 \;\leq\; L\frac{\rho_\nu}{\beta}\|d^\nu\|_M^2 \, ,$$

so that ρ_ν can be estimated by

$$\rho_\nu \geq -\beta\left(\frac{1-\delta}{L}\right)\frac{\psi'(0)}{\|d^\nu\|_M^2} \; .$$

Inserting this estimate into (5.31), we finally get

$$\psi(\rho_\nu) \leq \psi(0) - \delta\beta\left(\frac{1-\delta}{L}\right)\left(\frac{\psi'(0)}{\|d^\nu\|_M}\right)^2 \; .$$

□

Starting with $j = 0$, efficient step sizes can be computed from (5.29) by a finite number of tests. While the Armijo rule uses the values of h we will now present a step size rule using the gradient $\nabla h = H$ of h which guarantees efficient step sizes ρ_ν.

Proposition 5.4. *Let $d : \mathbb{N} \times \mathbb{R}^m \to \mathbb{R}^m$ generate descent directions. For a sequence $w^\nu \subset \mathbb{R}^m$ and directions $d^\nu = d(\nu, w^\nu)$ assume that a fixed parameter $\epsilon \in [0, 1)$ is given. Then any step rule ρ that satisfies $\rho_\nu = \rho(\nu, w^\nu, d^\nu) \geq 0$ and*

$$\left\langle \nabla h(w^\nu + \rho_\nu d^\nu), d^\nu \right\rangle \in [\epsilon \langle \nabla h(w^\nu), d^\nu \rangle, 0]$$

generates efficient step sizes that satisfy (5.18) *with*

$$c_S = \frac{1 - \epsilon^2}{2L}.$$

Proof. For $d^\nu = 0$ the assertion is obvious. Thus let $d^\nu \neq 0$ and hence $\psi'(0) \neq 0$. From $\epsilon\psi'(0) \leq \psi'(\rho_\nu)$ and Lipschitz continuity of ψ' we get

$$0 < \tilde{\rho}_\nu := \frac{(\epsilon - 1)}{L \|d^\nu\|_M^2} \psi'(0) \leq \rho_\nu.$$

Noting that ψ is decreasing in $[0, \rho_\nu]$ and using the definition of $\tilde{\rho}_\nu$ yields

$$\begin{aligned}
\psi(\rho_\nu) \leq \psi(\tilde{\rho}_\nu) &= \psi(0) + \int_0^{\tilde{\rho}_\nu} \psi'(t) - \psi'(0)\, dt + \tilde{\rho}_\nu \psi'(0) \\
&\leq \psi(0) + \tilde{\rho}_\nu \left(\frac{1}{2} L \|d^\nu\|_M^2 \tilde{\rho}_\nu + \psi'(0) \right) \\
&= \psi(0) + \tilde{\rho}_\nu \left(\left(\frac{\epsilon - 1}{2} \right) \psi'(0) + \psi'(0) \right) \\
&= \psi(0) + \tilde{\rho}_\nu \left(\frac{1 + \epsilon}{2} \right) \psi'(0) \\
&= \psi(0) - \left(\frac{1 - \epsilon}{L} \right) \left(\frac{1 + \epsilon}{2} \right) \left(\frac{\psi'(0)}{\|d^\nu\|_M} \right)^2.
\end{aligned}$$

□

Now we can obtain a sequence of efficient step sizes either by computing the first zero of ψ' exactly ($\epsilon = 0$) or by approximating it with fixed $0 \leq \epsilon < 1$. The latter can be done for example using the bisection method which requires one evaluation of ∇h per bisection step.

Observe that for both step size rules a sequence of evaluations of either h or ∇h is required. In view of Theorem 5.1 this will involve one solution of the minimization problem associated with F^{-1} per evaluation of h or ∇h, which can be very expensive. In order to mitigate this disadvantage it will be useful to decide a priori if we can choose $\rho_\nu = 1$ for a given w^ν and ν or if some kind of line search is needed. To this end let $\alpha_{-1} > 0$ and $\sigma \in (0,1)$ and define for $w^\nu, d^\nu \in \mathbb{R}^m$ the sequence

$$\alpha_\nu = \begin{cases} \|d^\nu\|_M & \text{if } \|d^\nu\|_M \leq \sigma\alpha_{\nu-1}, \\ \alpha_{\nu-1} & \text{else.} \end{cases} \tag{5.32}$$

For a step size rule ρ that generates efficient step sizes we will switch off the step rule if the norm of the direction decreases by the factor σ in the following sense:

$$\tilde{\rho}_\nu = \begin{cases} 1 & \text{if } \|d^\nu\|_M \leq \sigma\alpha_{\nu-1}, \\ \rho(\nu, w^\nu, d^\nu) & \text{else.} \end{cases} \tag{5.33}$$

Note that the sequence $\tilde{\rho}_\nu$ can easily be computed in practice. If $\|d^\nu\|_M \leq \sigma\alpha_{\nu-1}$ is not true, the step size $\tilde{\rho}_\nu$ is computed using the step size rule ρ. If the criterion is satisfied for some ν, the step size $\tilde{\rho}_{\nu'} = 1$ is used and the new bound $\alpha_\nu = \|d^\nu\|_M$ is computed. Thus the criterion for the ν-th step is checked with the bound $\alpha_{\nu-1} = \|d^{\nu'}\|_M$ where ν' is the last iteration step that satisfied the criterion.

It is also possible to simplify the criterion for the selection of $\tilde{\rho}_\nu = 1$ to the stronger criterion

$$\|d^\nu\|_M \leq \sigma \min_{\mu<\nu} \|d^\mu\|_M,$$

and the convergence proof of the following theorem remains essentially the same.

Theorem 5.4. *Assume that d and ρ generate gradient-related directions and efficient step sizes, respectively. Furthermore, assume that $d(\nu, v^\nu) \to 0$ implies $\nabla h(v^\nu) \to 0$ for any sequence v^ν. If $\tilde{\rho}_\nu$ is computed by (5.32) and (5.33) for some $\alpha_{-1} > 0$ and $\sigma \in (0,1)$ and $d^\nu = d(\nu, w^\nu)$, then the iterates w^ν obtained by*

$$\begin{aligned} w^{\nu+\frac{1}{2}} &= w^\nu + \tilde{\rho}_\nu d^\nu, \\ w^{\nu+1} &= w^{\nu+\frac{1}{2}} + \mathcal{C}(w^{\nu+\frac{1}{2}}), \end{aligned}$$

converge to w^* for an arbitrary initial iterate $w^0 \in \mathbb{R}^m$.

Proof. Define the set

$$\mathcal{N} = \left\{\nu \in \mathbb{N} \cup \{0\} \,:\, \|d^\nu\|_M \leq \sigma\alpha_{\nu-1}\right\}$$

of all iterates where the step size rule is not used. If $\mathcal{N}$ is bounded, the sequence w^ν converges by Theorem 5.2. From now on we assume that $\mathcal{N} = \{\eta_1 < \eta_2 < \eta_3 < \dots\}$ is unbounded.

For $\nu \in \mathcal{N}$ the monotonicity $h(w^{\nu+1}) \leq h(w^\nu)$ is not guaranteed because no line search is applied. Hence we have to show boundedness of $h(w^\nu)$ by other means. By monotonicity of $\mathcal{C}$, convexity of h, and Lipschitz continuity of ∇h we get

$$\begin{aligned} h(w^{\nu+1}) &\leq h(w^{\nu+\frac{1}{2}}) \\ &\leq h(w^{\nu}) - \left\langle \nabla h(w^{\nu+\frac{1}{2}}), w^{\nu} - w^{\nu+\frac{1}{2}} \right\rangle \\ &\leq h(w^{\nu}) + \left\langle \nabla h(w^{\nu}) - \nabla h(w^{\nu+\frac{1}{2}}), w^{\nu} - w^{\nu+\frac{1}{2}} \right\rangle - \left\langle \nabla h(w^{\nu}), w^{\nu} - w^{\nu+\frac{1}{2}} \right\rangle \\ &\leq h(w^{\nu}) + L\|w^{\nu} - w^{\nu+\frac{1}{2}}\|_M^2 + \tilde{\rho}_\nu \left\langle \nabla h(w^{\nu}), d^{\nu} \right\rangle \\ &\leq h(w^{\nu}) + L\tilde{\rho}_\nu^2 \|d^{\nu}\|_M^2 \end{aligned}$$

for all ν. Together with $\|d^{\eta_i}\|_M \leq \sigma^i \alpha_{-1}$ this implies

$$h(w^{\eta_i+1}) \leq h(w^{\eta_i}) + L\sigma^{2i}\alpha_{-1}^2$$

for $\eta_i \in \mathcal{N}$. Since we have $h(w^{\nu+1}) \leq h(w^{\nu})$ for all $\nu \notin \mathcal{N}$, we can apply the above estimate recursively yielding

$$\begin{aligned} h(w^{\eta_i+1}) &\leq h(w^{\eta_i}) + L\sigma^{2i}\alpha_{-1}^2 \\ &\leq h(w^{\eta_{i-1}+1}) + L\sigma^{2i}\alpha_{-1}^2 \\ &\leq h(w^{\eta_{i-1}}) + L\sigma^{2(i-1)}\alpha_{-1}^2 + L\sigma^{2i}\alpha_{-1}^2 \\ &\leq \cdots \leq h(w^{\nu_0}) + L\alpha_{-1}^2 \sum_{k=1}^{i} \sigma^{2k} \\ &\leq h(w^0) + \frac{L\alpha_{-1}^2}{1-\sigma^2} =: C. \end{aligned}$$

From $\|d^{\eta_i}\|_M \leq \sigma^i \alpha_{-1}$ we know that $d^{\eta_i} = d(\eta_i, d^{\eta_i}) \to 0$ and thus $\nabla h(w^{\eta_i}) \to 0$. Hence Lemma 5.2 implies $w^{\eta_i} \to w^*$ and $w^{\eta_i+\frac{1}{2}} = w^{\eta_i} + d^{\eta_i} \to w^*$.

Now let $w^{\nu_i} \to w^{**}$ be any convergent subsequence of w^{ν} with $\nu_0 > \eta_0$ and define

$$\tilde{\eta}_\nu = \max\{\eta \in \mathcal{N} \,:\, \eta < \nu\}.$$

Again using the monotonicity for $\nu \in \mathcal{N}$ we get

$$h(w^*) \leq h(w^{\nu_i}) \leq \cdots \leq h(w^{\tilde{\eta}_{\nu_i}+1}) \leq h(w^{\tilde{\eta}_{\nu_i}+\frac{1}{2}}) \to h(w^*)$$

yielding $h(w^{\nu_i}) \to h(w^*) = h(w^{**})$. Uniqueness of the minimizer implies $w^{**} = w^*$. Since this holds for any convergent subsequence, and there is at least one such subsequence (namely w^{η_i}) we have shown $w^{\nu} \to w^*$. □

We will see that an important example for directions satisfying the extra assumption of Theorem 5.4 is given by

$$d(\nu, w^{\nu}) = -S_\nu^{-1} \nabla h(w^{\nu})$$

with symmetric positive definite matrices S_ν that are bounded uniformly from above and below with respect to ν. If such directions are evaluated inexactly one does in general not know a priori if the inexact directions satisfy

$$\tilde{d}(\nu, v^\nu) \to 0 \qquad \Rightarrow \qquad \nabla h(w^\nu) \to 0.$$

In this case the following generalization of Theorem 5.4 can be used.

Corollary 5.2. *Let d and ρ satisfy the assumptions of Theorem 5.4 and let $\tilde{d}$ satisfy the assumptions of Proposition 5.2 with the accuracy condition* (5.24)*, i.e.,*

$$\|d^\nu - \tilde{d}^\nu\|_M \leq c\|\tilde{d}^\nu\|_M \quad \forall \nu \in \mathbb{N}.$$

If $\tilde{\rho}_\nu$ is computed by (5.32) *and* (5.33) *for some $\alpha_{-1} > 0$ and some $\sigma \in (0, 1)$ with d^ν replaced by $\tilde{d}^\nu = \tilde{d}(\nu, w^\nu)$, then the iterates w^ν obtained by*

$$\begin{aligned} w^{\nu+\frac{1}{2}} &= w^\nu + \tilde{\rho}_\nu \tilde{d}^\nu, \\ w^{\nu+1} &= w^{\nu+\frac{1}{2}} + \mathcal{C}(w^{\nu+\frac{1}{2}}), \end{aligned}$$

converge to w^* for an arbitrary initial iterate $w^0 \in \mathbb{R}^m$.

Proof. By Proposition 5.2 the directions generated by $\tilde{d}$ are also gradient related. Furthermore, the accuracy condition (5.24) implies

$$\|d^\nu\|_M \leq \|d^\nu - \tilde{d}^\nu\|_M + \|\tilde{d}^\nu\|_M \leq (1 + c)\|\tilde{d}^\nu\|_M.$$

Hence $\tilde{d}(\nu, v^\nu) \to 0$ implies $d(\nu, v^\nu) \to 0$ and thus $\nabla h(w^\nu)$. Now Theorem 5.4 can be applied with d replaced by $\tilde{d}$. □

The above result does no longer hold if $\tilde{d}$ satisfies the assumptions of Proposition 5.2 with the accuracy condition (5.25) instead of (5.24), i.e.,

$$\|d^\nu - \tilde{d}^\nu\|_M \leq c\|d^\nu\|_M \quad \forall \nu \in \mathbb{N}.$$

In this case $\tilde{d}(\nu, v^\nu) \to 0$ does not imply $d(\nu, v^\nu) \to 0$. Hence we do no longer know if $\nabla h(w^\nu) \to 0$ is implied.

5.3 Derivatives of the Nonlinear Schur Complement

The convergence speed of gradient-related descent algorithms depends heavily on the selection of the descent directions d^ν. If h is C^2 the directions

$$d^\nu = -(\nabla^2 h(w^\nu))^{-1} \nabla h(w^\nu) \tag{5.34}$$

lead to a damped Newton method for the operator $H = \nabla h$. If H is not differentiable but Lipschitz continuous we want to define directions similar to (5.34), replacing

$(\nabla^2 h(w^\nu))$ by symmetric positive definite matrices $S(w^\nu) \in \mathbb{R}^{m,m}$ that represent generalized linearizations of H at w^ν.

If not stated otherwise we assume for the rest of this chapter that (A5), (A8), (A10) and (A11) are satisfied and that h and $H = \nabla h$ are given as in Theorem 5.1. For the special case of a quadratic obstacle problem with additional linear constraint such linearizations $S(w^\nu)$ were introduced by Gräser and Kornhuber [59]. There the piecewise linearity of H in that case was used. Here we will generalize this approach using piecewise smoothness of H instead.

5.3.1 Derivatives of F^{-1}

Since H is a composition and sum of affine functions with F^{-1} the crucial part in the derivation of linearizations of H are linearizations of F^{-1}. In order to derive such linearizations for F^{-1} we first look at the functionals φ_i. Again it will be helpful to consider the limits $\varphi'_{i,-}$,$\varphi'_{i,+}$,$\varphi''_{i,-}$,$\varphi''_{i,+}$ from Lemma 4.2, which are essentially one-sided first and second derivatives. In view of Lemma 4.3, φ''_i will again denote one of the one-sided second derivatives if they do not coincide, and ∞ if one of them is ∞.

As in the case of the nonlinear smoothers the linearization of F^{-1} will be defined piecewise and the components i where φ_i lacks regularity need special care. Thus we again use the inactive sets

$$\mathcal{I}(v) := \{i : \partial\varphi_i(v_i) \text{ is single-valued}\}$$

defined in (4.24). For convenience we also define the corresponding active sets

$$\mathcal{A}(v) := \{1, \dots, n\} \setminus \mathcal{I}(v).$$

In order to extract a decomposition of $\mathbb{R}^n$ into nontrivial subsets where F^{-1} is smooth we have to distinguish different active configurations. Since the functions φ_i may have multiple points a_i^k where they are not smooth, an active configuration is not completely determined by the active set itself. To distinguish different configurations we also have to take the values at the active component into account. The equivalence classes

$$[c] := \{v \in \operatorname{dom}\varphi : \mathcal{A}(v) = \mathcal{A}(c), v_i = c_i \, \forall i \in \mathcal{A}(c)\}$$

defined for $c \in \operatorname{dom}\varphi$ containing all vectors with the same active configuration provide exactly this distinction. Hence we can address an active configuration by $[c]$ for one representative. By definition x and y have the same active configuration if and only if $[x] = [y]$ and hence the representative is obviously not unique. The set of all possible active configurations is given by

$$\mathbb{A} := \{[c] : c \in \operatorname{dom}\varphi\}.$$

By Assumption (A8) the set $\mathbb{A}$ is finite and $\operatorname{dom}\varphi$ can be decomposed according to

$$\operatorname{dom}\varphi = \bigcup_{[c]\in\mathbb{A}} [c].$$

Since F has a single-valued inverse we have $\operatorname{dom}\varphi \supset F^{-1}(\mathbb{R}^n)$ and thus $\mathbb{R}^n = F(\operatorname{dom}\varphi)$ can be decomposed according to

$$\mathbb{R}^n = \bigcup_{[c]\in\mathbb{A}} F([c]), \qquad F([c]) = \bigcup_{x\in[c]} F(x) = \{y : F^{-1}(y) \in [c]\}$$

using the images of $[c]$ under F. Note that in general $\operatorname{dom}\varphi \supset F^{-1}(\mathbb{R}^n)$ is a real inclusion and equality does not hold. This is due to the possibility of $\varphi_i'(x_i) \to \infty$ for $x_i \to a_i^{m_i}$. In this case $F(x)$ is even empty for all $x \in \operatorname{dom}\varphi$ with $x_i = a_i^{m_i}$.

We will define the linearization of F^{-1} piecewise on sets where the operator is smooth. Thus we do not only need to handle the active components but also the smoothness intervals (a_i^{k-1}, a_i^k) the inactive components are contained in. To this end it is convenient to first define the set $\mathcal{E}$ of all multi-indices needed to identify these intervals by

$$\mathcal{E} := \{\eta \in \mathbb{N}^n : 1 \le \eta_i \le m_i\}.$$

Now we can define the sets of all vectors corresponding to an active configuration $[c] \in \mathbb{A}$ and the open and closed smoothness intervals $\eta \in \mathcal{E}$ by

$$\begin{aligned} [c]_\eta &:= \{x \in [c] : x_i \in (a_i^{\eta_i - 1}, a_i^{\eta_i}) \text{ for } i \in \mathcal{I}(c)\}, \\ [[c]]_\eta &:= \{x \in [c] : x_i \in [a_i^{\eta_i - 1}, a_i^{\eta_i}] \text{ for } i \in \mathcal{I}(c)\}. \end{aligned}$$

Both sets are $(n - |\mathcal{I}(c)|)$-dimensional hypercubes. The set $[[c]]_\eta$ is in general only a subset of $\overline{[c]}_\eta$ since $\operatorname{dom}\varphi_i$ may not contain a_i^0 and $a_i^{m_i}$. These sets provide a decomposition of $[c]$ in the sense that

$$[c] = \bigcup_{\eta\in\mathcal{E}} [[c]]_\eta, \qquad \overline{[c]} = \bigcup_{\eta\in\mathcal{E}} \overline{[c]}_\eta.$$

Note that these decompositions are not disjoint in general.

If a linearization of F^{-1} is to be defined piecewise it is important that the sets where it is defined do not degenerate to lower-dimensional objects or, equivalently, that active configurations are stable in a certain sense. This is provided by the following lemma.

Lemma 5.4. *Let $[c] \in \mathbb{A}$ with $F(c) \neq \emptyset$ and $\eta \in \mathcal{E}$. Then*

$$F([[c]]_\eta) \subset \overline{F([c]_\eta)^\circ} \subset \overline{\operatorname{int} F([[c]]_\eta)} \neq \emptyset$$

holds for the open set

$$F([c]_\eta)^\circ := \left\{ y \in F([c]_\eta) : \left\{ \begin{array}{ll} y \in (\nabla J_0(F^{-1}(y)))_i + \operatorname{int}\partial\varphi_i(F^{-1}(y)_i) & \forall i \in \mathcal{A}(c), \\ F^{-1}(y)_i \in (a_i^{\eta_i - 1}, a_i^{\eta_i}) & \forall i \in \mathcal{I}(c) \end{array} \right. \right\}.$$

Proof. Let $[c] \in \mathbb{A}$ with $F(c) \neq \emptyset$ and $\eta \in \mathcal{E}$. Since $\partial\varphi_i(c_i)$ is set-valued for $i \in \mathcal{A}(c)$, an element of $F([c]_\eta)^\circ$ can easily be constructed, which shows that $F([c]_\eta)^\circ \neq \emptyset$. Next we show that $F([c]_\eta)^\circ$ is open.

Let $y \in F([c]_\eta)^\circ$, $x = F^{-1}(y)$ be fixed and $x' = F^{-1}(y')$ for some y' with $\|y-y'\|_\infty < \epsilon$. By (A8) and continuity of F^{-1} we instantly get

$$x'_i \in (a_i^{\eta_i-1}, a_i^{\eta_i})$$

and thus $\mathcal{A}(x') \subset \mathcal{A}(x)$ if ϵ is small enough.

To show $\mathcal{A}(x') \supset \mathcal{A}(x)$ assume that $x'_{\mathcal{I}(x)}$ is known and fixed. Then $x'_{\mathcal{A}(x)}$ is the unique solution of

$$F(x'_{\mathcal{I}(x)} + x'_{\mathcal{A}(x)})_i \ni y'_i \qquad \forall i \in \mathcal{A}(x). \tag{5.35}$$

By continuity of F^{-1} and ∇J_0 the residual defined by $r(b,v) := b - \nabla J_0(v)$ satisfies

$$\Big\| r(y,x) - r(y', x'_{\mathcal{I}(x)} + x_{\mathcal{A}(x)}) \Big\|_\infty < \max\Big\{ \operatorname{dist}\big(\partial P_{x,i}, r(y,x)_i\big) : i \in \mathcal{A}(x) \Big\}$$

for the border $\partial P_{x,i}$ of the set $P_{x,i} = \partial\varphi_i(x_i)$ if ϵ is small enough. In this case we have $r(y', x'_{\mathcal{I}(x)} + x_{\mathcal{A}(x)}) \in \operatorname{int} \partial\varphi_i(x_i)$. Hence $x'_{\mathcal{A}(x)} = x_{\mathcal{A}(x)}$ solves (5.35) which yields $\mathcal{A}(x') = \mathcal{A}(x)$. Thus $x' \in [c]_\eta$ and even more $y' \in F([c]_\eta)^\circ$. Since y was arbitrary, $F([c]_\eta)^\circ$ must be open and we have $F([c]_\eta)^\circ \subset \operatorname{int} F([[c]]_\eta)$.

Now let $y \in F([[c]]_\eta) \setminus F([c])^\circ_\eta$ with $x = F^{-1}(y)$ be fixed. Then it is easy to give a sequence $x^k \in [c]_\eta$ with $x^k \to x$. For the sequence

$$y^k = \nabla J_0(x^k) + (y - \nabla J_0(x) + z^k)_{\mathcal{A}(c)} + (\partial\varphi(x^k))_{\mathcal{I}(c)}$$

with

$$z_i^k = \frac{\epsilon}{k} \begin{cases} 1 & \text{if } i \in \mathcal{A}(c) \text{ and } (y - \nabla J_0(x))_i = \min \partial\varphi_i(c_i), \\ -1 & \text{if } i \in \mathcal{A}(c) \text{ and } (y - \nabla J_0(x))_i = \max \partial\varphi_i(c_i), \\ 0 & \text{else} \end{cases}$$

and ϵ small enough we have

$$(y^k - \nabla J_0(x^k))_i = \begin{cases} (y - \nabla J_0(x) + z^k)_i \in \operatorname{int} \partial\varphi_i(c_i) & \text{if } i \in \mathcal{A}(c), \\ \partial\varphi_i(x^k) & \text{if } i \in \mathcal{I}(c) \end{cases}$$

and hence $x^k = F^{-1}(y^k)$ and $y^k \in F([c])^\circ_\eta$. Since ∇J_0 is continuous and φ_i is continuously differentiable on $(a_i^{\eta_i-1}, a_i^{\eta_i+1})$ for $i \in \mathcal{I}(c)$ we have

$$y^k \to \nabla J_0(x) + (y - \nabla J_0(x))_{\mathcal{A}(c)} + (\partial\varphi(x))_{\mathcal{I}(c)} = y,$$

which proves the assertion. □

Since $[c]$ decomposes into the sets $[[c]]_\eta$, Lemma 5.4 implies

$$F([c]) \subset \overline{\operatorname{int} F([c])}.$$

While this lemma shows that the sets $F([c])$ do not degenerate it does not give insight into their structure. The following remark sheds some light on the geometry of these sets.

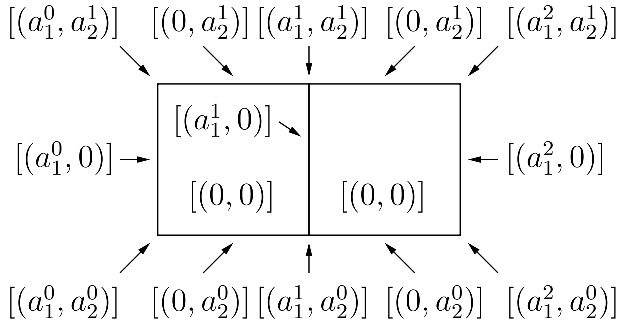


Figure 5.2: Decomposition of $\operatorname{dom}\varphi \subset \mathbb{R}^2$ into the sets $[c]$, $c \in \mathbb{A} \subset \mathbb{R}^2$.

Remark 5.2. *Define the points where φ_i is not differentiable by*

$$\{\tilde{a}_i^0, \dots, \tilde{a}_i^{\widetilde{m}_i}\} := \{a : \partial\varphi_i(a) \text{ is set-valued}\} \subset \{a_i^0, \dots, a_i^{m_i}\}.$$

Then the configuration $[c']$ without active component, i.e. $\mathcal{A}(c') = \emptyset$, is clearly given by the open set

$$[c'] = \prod_{i=1}^{n} \bigcup_{k=1}^{\widetilde{m}_i} (\tilde{a}_i^{k-1}, \tilde{a}_i^k) = \bigcup_{\substack{(k_1,\dots,k_n) \\ 1 \le k_i \le \widetilde{m}_i}} \prod_{i=1}^{n} (\tilde{a}_i^{k_i - 1}, \tilde{a}_i^{k_i})$$

and a representative is, e.g., given by $c_i' = \frac{1}{2}(\tilde{a}_i^0 + \tilde{a}_i^1)$. Note that $[c']$ is the union of n-dimensional open hypercubes $Q_{(k_1,\dots,k_n)}$. If the arguments of Lemma 5.4 are applied with the indicator functions of these hypercubes instead of φ it can be seen that $\nabla J_0(Q_{(k_1,\dots,k_n)}) \subset \overline{\operatorname{int} \nabla J_0(Q_{(k_1,\dots,k_n)})}$. Hence the images of the hypercubes under ∇J_0 do not degenerate in the sense that all points are limits of sequences in their interior.

If at least one component of c is active the set $[c]$ is the union of hypercubes Q with dimension less then n, and hence no longer open in $\mathbb{R}^n$. To be precise the length of these hypercubes in any direction e_i with $i \in \mathcal{A}(c)$ is zero. However, the set $(\partial\varphi([c]))_{\mathcal{A}(c)}$ is a hypercube that has nonzero lengths exactly in the directions e_i with $i \in \mathcal{A}(c)$.

Figure 5.2 and Figure 5.3 show an example of the decomposition of $\operatorname{dom}\varphi$ and $\mathbb{R}^2$ into the sets $[c] \in \mathbb{A}$ and $F([c])$, respectively. For simplicity it is assumed that all a_i^k differ from 0, such that $c_i = 0$ means that the i-th component is not active. While the sets $[c]$ are 1-dimensional edges or 0-dimensional vertices if one or two components of c are active, the corresponding images $F([c])$ of all such sets have a nontrivial interior. Note that $F([c])$ has a curved boundary in general but edges parallel to the i-th axis if $c_i = \tilde{a}_i^k$ for some k. For example the set $F([(0,0)])$ of all $F(x)$ such that φ_i is smooth at x_i for all i might have all edges curved. Conversely, the set $F([(a_1^1, 0)])$ of all $F(x)$ such that the first component is fixed to the kink a_1^1 (and thus active) has two straight edges parallel to the first axis. In case of a quadratic function J_0 all $F([c])$, $[c] \in \mathbb{A}$, are (possibly unbounded) parallelepipeds.

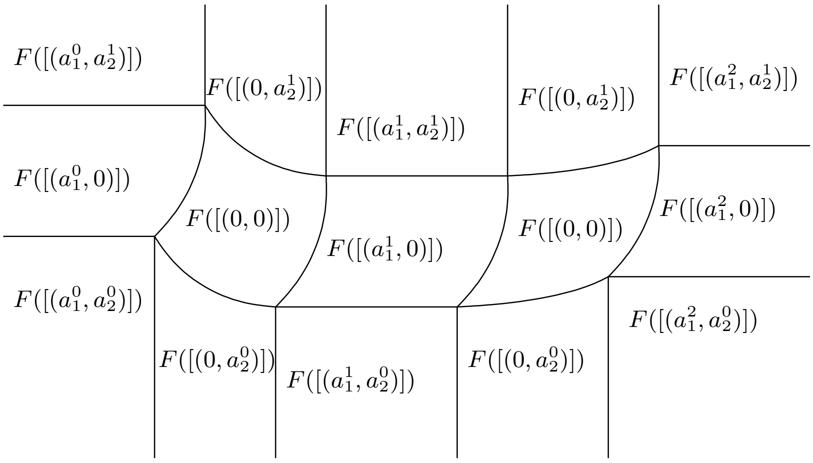


Figure 5.3: Decomposition of $\mathbb{R}^2$ into the sets $F([c])$, $c \in \mathbb{A} \subset \mathbb{R}^2$.

In the following we will derive a generalized linearization of F^{-1} that is defined piecewise on the sets $F([[c]]_\eta)$. In order to do this we first investigate these sets further by rewriting F in terms of a Lipschitz continuous and a diagonal operator.

Lemma 5.5. *Let $x \in \operatorname{dom} \varphi$. Then F can by represented by*

$$F(x) = \underbrace{\left[(\nabla J_0 - I) \circ (I + \partial\varphi)^{-1} + I\right]}_{=:T} \left((I + \partial\varphi)(x)\right).$$

The operator T is Lipschitz continuous.

Proof. Ix is the gradient of the convex functional $\frac{1}{2}\| \cdot \|^2$ at x. By Proposition 4.1 the operator $I + \partial\varphi$ has a single-valued Lipschitz continuous inverse which yields the representation of $F(x)$. By Assumption (A3) the operator ∇J_0 and thus T is Lipschitz. □

Lemma 5.6. *Let $[c] \in \mathbb{A}$ with $F(c) \neq \emptyset$ and $\eta \in \mathcal{E}$. Then*

$$(I + \partial\varphi)([[c]]_\eta) \subset \overline{Q(c, \eta)}$$

with the nonempty open hypercubes $Q(c, \eta) = \prod_{i=1}^n Q(c, \eta, i)$ spanned by the open intervals

$$Q(c, \eta, i) = \begin{cases} c_i + \operatorname{int} \partial\varphi_i(c_i) & \text{if } i \in \mathcal{A}(c), \\ (a_i^{\eta_i - 1} + \varphi'_{i,+}(a_i^{\eta_i - 1}), a_i^{\eta_i} + \varphi'_{i,-}(a_i^{\eta_i})) & \text{if } i \in \mathcal{I}(c). \end{cases}$$

Proof. We only have to note that $I + \partial\varphi$ is strictly and maximal monotone and that $a_i^{\eta_i - 1} < a_i^{\eta_i}$. □

Lemma 5.7. *Let $[c] \in \mathbb{A}$ with $F(c) \neq \emptyset$ and $\eta \in \mathcal{E}$. Then $F([[c]]_\eta) \setminus F([c]_\eta)^\circ$ is a set of measure zero.*

Proof. For $i \in \mathcal{A}(x)$ the set $\partial\varphi_i(x_i)$ is set-valued and convex. Hence its interior is not empty. This and the fact that $(1 + \partial\varphi_i(x_i))^{-1}$ is single-valued implies

$$x_i = (1 + \partial\varphi_i)^{-1}(1 + \partial\varphi)(x_i) = (1 + \partial\varphi_i)^{-1}(1 + \operatorname{int}\partial\varphi)(x_i).$$

Thus we get $I = (I + \partial\varphi)^{-1}(I + \widetilde{\partial\varphi})$ for

$$\widetilde{\partial\varphi}_i(x_i) = \begin{cases} \operatorname{int}\partial\varphi_i(x_i) & \text{if } i \in \mathcal{A}(c), \\ \partial\varphi_i(x_i) & \text{if } i \in \mathcal{I}(c). \end{cases}$$

For $y \in F([[c]]_\eta)$, $x = F^{-1}(y)$, and $i \in \mathcal{A}(c)$ we especially get

$$y_i \in \nabla J_0(x)_i + \operatorname{int}\partial\varphi_i(x_i) \quad \Leftrightarrow \quad y_i \in T(I + \widetilde{\partial\varphi})(x)$$

with the operator T of Lemma 5.5. Using this and Lemma 5.6 we get

$$F([c]_\eta)^\circ = T(Q(c,\eta)), \qquad F([[c]]_\eta) = T(I + \partial\varphi)([[c]]_\eta) \subset T(\overline{Q(c,\eta)})$$

and thus

$$F([[c]]_\eta) \setminus F([c]_\eta)^\circ \subset T(\partial Q(c,\eta)). \tag{5.36}$$

Since $\partial Q(c,\eta)$ is the boundary of an open hypercube, its measure is zero. Together with the Lipschitz continuity of T this yields that $T(\partial Q(c,\eta))$ and thus $F([[c]]_\eta) \setminus F([c]_\eta)^\circ$ has also measure zero (see, e.g., [110, Lemma 2.3]). □

The decomposition of $\mathbb{R}^m$ suggests to define linearizations of F^{-1} on each set $F([c]_\eta)$ or $F([[c]]_\eta)$ separately. However, this is not completely straightforward due to possibly unbounded derivatives at the boundaries of these sets. In view of Lemma 5.4 we thus first derive linearizations on the open subsets $F([c]_\eta)^\circ$.

Consider the operator F^{-1} on the set $F([[c]]_\eta)$ for some fixed $[c] \in \mathbb{A}$ with $F(c) \neq \emptyset$ and $\eta \in \mathcal{E}$. On this set $x = F^{-1}y$ is equivalent to

$$x_{\mathcal{A}(c)} = c_{\mathcal{A}(c)} \qquad \text{and} \qquad \underbrace{F(c_{\mathcal{A}(c)} + x_{\mathcal{I}(c)})_{\mathcal{I}(c)}}_{=:G_c(x_{\mathcal{I}(c)})} = y_{\mathcal{I}(c)}. \tag{5.37}$$

Note that this is a real equation and not only an inclusion because the restriction of F to the components $i \in \mathcal{I}$ is single-valued. The operator G_c is the restriction of F to the inactive components in the configuration $[[c]]_\eta$. Equation (5.37) for $x_{\mathcal{I}(c)}$ is equivalent to the restriction of the minimization problem associated with F to these components. Since the smooth part of G_c, which incorporates ∇J_0, also satisfies the strong convexity (4.3) for $u, v \in V_{\mathcal{I}(c)} = \{x \in \mathbb{R}^n : x = x_{\mathcal{I}(c)}\}$, the operator

$$G_c : (\overline{[[c]]}_\eta - c_{\mathcal{A}(c)}) \to F(\overline{[[c]]}_\eta) \cap V_{\mathcal{I}(c)}$$

is invertible and F^{-1} can be written as

$$F^{-1}(y) = G_c^{-1}(y_{\mathcal{I}(c)}) + c_{\mathcal{A}(c)}. \tag{5.38}$$

By definition of the set $F([c]_\eta)^\circ$ we know that φ_i is twice continuously differentiable at $(F^{-1}(y))_i$ for $y \in F([c]_\eta)^\circ$ and $i \in \mathcal{I}(c)$. Thus the restriction of G^{-1} on the set $F([c]_\eta)^\circ$ is differentiable. Having this representation it is also clear that for $i \in \mathcal{A}(c)$ the i-th row and column of a linearization of F^{-1} should be zero, since $F^{-1}(y)$ is constants in the i-th component and does not depend on the i-th component of its argument y. The remaining entries are given by the derivative of G_c^{-1}.

The following result for the case of a differentiable ∇J_0 does even incorporate the case $\varphi_i' \to \pm\infty$ that was excluded until now by only considering the open subset $F([c]_\eta)^\circ$.

Theorem 5.5. *Let J_0 be twice continuously differentiable. Then an element of the generalized derivative in the sense of Clarke at $y = F(x)$ is given by the pseudoinverse $(\partial^2 J(x)_{\mathcal{I}'(x)})^+$ of the reduced Hessian $\partial^2 J(x)_{\mathcal{I}'(x)}$ with the reduced inactive set $\mathcal{I}'(x)$ as defined in* (4.25) *and* (4.34)*, respectively. I.e., we have*

$$\left(\partial^2 J(x)_{\mathcal{I}'(x)}\right)^+ \in \partial_B(F^{-1})(y) \subset \partial_C(F^{-1})(y). \tag{5.39}$$

Furthermore, F^{-1} is differentiable on each set $F([c]_\eta)^\circ$ with $[c] \in \mathbb{A}$ and $\eta \in \mathcal{E}$ and the derivative is given by the matrix $(\partial^2 J(x)_{\mathcal{I}'(x)})^+$.

Proof. First we recall that the matrix $\partial^2 J(x)_{\mathcal{I}'(x)}$ defined in (4.25) and (4.34) is well-defined since for $i \in \mathcal{I}'(x)$ the function φ_i is either twice continuously differentiable at x_i or the left and right one-sided second derivatives are finite.

Let $[c] \in \mathbb{A}$, with $F(c) \neq \emptyset$ and $\eta \in \mathcal{E}$ be fixed and consider $y \in F([c]_\eta)^\circ$ and $x = F^{-1}(y)$. In view of (5.38) we investigate the operator

$$G_c(x) = I_{\mathcal{I}(c)} F(c_{\mathcal{A}(c)} + I_{\mathcal{I}(c)} x).$$

G_c is differentiable on $[c]_\eta$ and its classical derivative is given by

$$\nabla G_c(x) = \left(\nabla^2 J_0(x_{\mathcal{I}(c)} + c_{\mathcal{A}(c)}) + \varphi''(x_{\mathcal{I}(c)})\right)_{\mathcal{I}(c)}.$$

Hence $G_c^{-1} : F([c]_\eta)^\circ \to [c]_\eta$ is also smooth. If we consider $\nabla G_c(x)$ as an operator from $V_{\mathcal{I}(c)}$ to $V_{\mathcal{I}(c)}$ it is invertible and the matrix representing the inverse is given by

$$\left(\nabla^2 J_0(x)_{\mathcal{I}'(x)} + \varphi''(x)_{\mathcal{I}'(x)}\right)^+ \tag{5.40}$$

where we have used that $\mathcal{I}'(x) = \mathcal{I}(x) = \mathcal{I}(c)$ due to $x \in [c]_\eta \subset [c]$. Since F^{-1} and G_c^{-1} differ only by a constant F^{-1} is differentiable at y and the derivative is also given by (5.40).

Now let $y \in F([c])$ with $x = F^{-1}(y)$ such that y is not contained in any of the sets $F([c]_\eta)^\circ$. Then $y \in F([[c]]_\eta) \setminus F([c]_\eta)^\circ$ is still true for some η and by definition φ_i is once but not twice differentiable at x_i for $i \in \mathcal{I}(c)$. Without loss of generality assume that η is chosen such that for $i \in \mathcal{I}(c) = \mathcal{I}(x)$

$$\xi^k \in (a_i^{\eta_i - 1}, a_i^{\eta_i}), \xi^k \to x_i \qquad \Rightarrow \qquad \varphi_i''(\xi^k) \to \begin{cases} \varphi_i''(x_i) & \text{if } i \in \mathcal{I}'(x), \\ \infty & \text{if } i \in \mathcal{I}(x) \setminus \mathcal{I}'(x) \end{cases}$$

holds true where $\varphi_i''(x_i)$ is the one-sided derivative selected in (4.34). (Otherwise choose the appropriate η' by increasing or decreasing the corresponding indices η_i by one.)

As in the proof of Lemma 5.4 let (x^k) be a sequence with $x^k \in [c]_\eta$ and $x^k \to x$. By the choice of η and the continuity of $\varphi_i'', i \in \mathcal{I}(x)$ on $[c]_\eta$ this sequence can in particular be chosen such that

$$\varphi_i''(x_i^k) = \alpha^k, \qquad \forall i \in \mathcal{I}(x) \setminus \mathcal{I}'(x)$$

for a fixed constant $\alpha > 1$. For this sequence construct $y^k \in F([c]_\eta)^\circ$ with $x^k = F^{-1}(y^k)$ and $y^k \to y$ as in the proof of Lemma 5.4 and define the sequence of matrices M^k by

$$M^k := \nabla^2 J_0(x^k)_{\mathcal{I}(x)} + \varphi''(x^k)_{\mathcal{I}'(x)} + I - I_{\mathcal{I}(x)}.$$

Then we have

$$M^k \to M := \nabla^2 J_0(x)_{\mathcal{I}(x)} + \varphi''(x)_{\mathcal{I}'(x)} + I - I_{\mathcal{I}(x)}.$$

Application of Lemma A.5 and Lemma A.6 in the appendix together with

$$\mathcal{I}'(x^k) = \mathcal{I}(x^k) = \mathcal{I}(x)$$

yields

$$\begin{aligned} \lim_{k\to\infty} \left[\nabla^2 J_0(x^k) + \varphi''(x^k)\right]^+_{\mathcal{I}'(x^k)} &= \lim_{k\to\infty} \left[\nabla^2 J_0(x^k) + \varphi''(x^k)\right]^+_{\mathcal{I}(x)} \\ &= \left[\lim_{k\to\infty} \left(M^k + \alpha^k I_{\mathcal{I}(x)\setminus\mathcal{I}'(x)}\right)^{-1}\right]_{\mathcal{I}(x)} \\ &= \left[\left(M_{\mathcal{A}(x)\cup\mathcal{I}'(x)}\right)^+\right]_{\mathcal{I}(x)} \\ &= \left(\nabla^2 J_0(x) + \varphi''(x)\right)^+_{\mathcal{I}'(x)}. \end{aligned}$$

This proves the assertion. □

Note that the case $|\varphi_i'(\xi^k)| \to \infty$ for $\xi^k \to x_i$ is included in Theorem 5.5. It does not need any special treatment since $\partial\varphi_i(x_i)$ and thus $F(x)$ are empty in this case.

Theorem 5.6. *Let J_0 be twice continuously differentiable and let all φ_i'' be uniformly bounded from above by a constant $c_{\varphi''}$ where $\partial\varphi_i$ is single-valued. Then there is a constant $c > 0$ such that F^{-1} is strongly monotone with respect to the semi-norm introduced by $cI_{\mathcal{I}_0}$, i.e.*

$$\left\langle F^{-1}(u) - F^{-1}(v), u - v\right\rangle \geq c \left\langle u - v, u - v\right\rangle_{\mathcal{I}_0} \qquad \forall u, v \in \mathbb{R}^n, \qquad (5.41)$$

where $\mathcal{I}_0$ is the smallest inactive set, i.e.

$$\mathcal{I}_0 := \bigcap_{y \in \mathbb{R}^n} \mathcal{I}'(F^{-1}(y)) = \mathbb{N} \setminus \{i \in \mathbb{N} : \exists \xi \in \mathbb{R} : \partial\varphi_i(\xi) \text{ is set-valued}\}\,.$$

Proof. By Theorem 5.5 F^{-1} is differentiable on the open set

$$\bigcup_{[c]\in\mathbb{A},\eta\in\mathcal{E}} F([c]_\eta)^\circ \subset \mathcal{D}_{F^{-1}}$$

and by Lemma 5.7

$$\mathcal{S} := \mathcal{D}_{F^{-1}} \setminus \bigcup_{[c]\in\mathbb{A},\eta\in\mathcal{E}} F([c]_\eta)^\circ \subset \bigcup_{[c]\in\mathbb{A},\eta\in\mathcal{E}} F([[c]]_\eta) \setminus F([c]_\eta)^\circ$$

has measure zero. Since φ_i'' is bounded from above we have

$$\left\langle [\nabla^2 J_0(x) + \varphi''(x)]_{\mathcal{I}'(x)} v_{\mathcal{I}'(x)}, v_{\mathcal{I}'(x)} \right\rangle \leq (\lambda_{\max}(\overline{H}_{J_0}) + c_{\varphi''}) \left\langle v_{\mathcal{I}'(x)}, v_{\mathcal{I}'(x)} \right\rangle$$

and thus the derivative of F^{-1} satisfies

$$\begin{aligned}\left\langle \left(\nabla^2 J_0(x) + \varphi''(x)\right)^+_{\mathcal{I}'(x)} v, v \right\rangle &= \left\langle \left(\nabla^2 J_0(x) + \varphi''(x)\right)^+_{\mathcal{I}'(x)} v_{\mathcal{I}'(x)}, v_{\mathcal{I}'(x)} \right\rangle \\ &\geq (\lambda_{\max}(\overline{H}_{J_0}) + c_{\varphi''})^{-1} \left\langle v_{\mathcal{I}'(x)}, v_{\mathcal{I}'(x)} \right\rangle \\ &\geq (\lambda_{\max}(\overline{H}_{J_0}) + c_{\varphi''})^{-1} \left\langle v_{\mathcal{I}_0}, v_{\mathcal{I}_0} \right\rangle\end{aligned}$$

for $y \in \mathcal{D}_{F^{-1}} \setminus \mathcal{S}$ and $x = F^{-1}(y)$. Now the application of Lemma A.3 in the appendix to $T = F^{-1}$ yields (5.41). □

In the more general case ∇J_0 is not differentiable. However, some kind of linearization $\partial^2 J_0(x)$ of ∇J_0 at $x = F^{-1}(y)$ is given by (A4). Hence we will use

$$\left(\partial^2 J(x)_{\mathcal{I}'(x)}\right)^+ = \left(\partial^2 J_0(x) + \varphi''(x)\right)^+_{\mathcal{I}'(x)}$$

as defined in (4.34) as linearization of F^{-1} at y in this case also.

5.3.2 Derivatives of H

If F^{-1} is a continuously differentiable operator we can easily derive a linearization of the nonlinear Schur complement

$$H(w) = -BF^{-1}(f - B^T w) + Cw + g$$

using the chain rule. The result is

$$\nabla H(w) = B\nabla(F^{-1})(f - B^T w)B^T + C.$$

If F itself is also differentiable we have $\nabla(F^{-1})(y) = (\nabla F)(F^{-1}(y))^{-1}$ and $\nabla H(w)$ as given above is just the Schur complement of the linear saddle point problem

$$u \in \mathbb{R}^n, w \in \mathbb{R}^m : \qquad \begin{pmatrix} (\nabla F)(F^{-1}(f - B^T w_0)) & B^T \\ B & -C \end{pmatrix} \begin{pmatrix} u \\ w \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix},$$

which is the linearization of the nonlinear saddle point problem (5.1) at $(u_0, w_0)^T$ with $u_0 = F^{-1}(f - B^T(w_0))$.

In the general case these derivatives do not exist. While F is not even a single-valued operator we know from Propositions 4.1 and 5.1 that F^{-1} and H are Lipschitz continuous. Thus one could in principle select elements of the generalized Jacobian

$$\partial_C H(w) = \operatorname{co} \partial_B H(w).$$

However, it will be complicated to compute elements of this set since the generalized Jacobian ∂_C does not satisfy the chain rule in general. Nevertheless we use a chain rule to obtain a generalized linearization $S(w)$ of H at w which is not necessarily an element of $\partial_C H(w)$. Based on the linearization of F^{-1} derived in the previous subsection this approach results in

$$S(w) := B\Big(\partial^2 J(u)_{\mathcal{I}'(u)}\Big)^+ B^T + C$$

as linearization of H at w with $u = F^{-1}(f - B^T w)$.

Proposition 5.5. *Let J_0 be twice continuously differentiable and let* $\operatorname{rank} B = n$. *Then*

$$S(w) \in \partial_B H(w) \subset \partial_C H(w) \qquad \forall w \in \mathbb{R}^m.$$

Proof. If $\operatorname{rank} B = n$ the mapping defined by $G(w) = f - B^T w$ is surjective. In the proof of Theorem 5.5 the generalized derivative of F^{-1} was derived as a limit of classical derivatives that are defined on disjoint open sets $F([c]_\eta)^\circ$ where F^{-1} is differentiable. Furthermore, the space $\mathbb{R}^n$ can be decomposed according to

$$\mathbb{R}^n = \bigcup_{[c]\in\mathbb{A},\eta\in\mathcal{E}} \overline{F([c]_\eta)^\circ} = \overline{\bigcup_{[c]\in\mathbb{A},\eta\in\mathcal{E}} F([c]_\eta)^\circ}.$$

Since F^{-1} is differentiable on $F([c]_\eta)^\circ$ this is also true for $F^{-1}\circ G$ and H on $G^{-1}(F([c]_\eta)^\circ)$. By the classical chain rule we have $\nabla H(w) = S(w)$ at $w \in G^{-1}(F([c]_\eta)^\circ)$. Now let

$$w \in R := \overline{\bigcup_{[c]\in\mathbb{A},\eta\in\mathcal{E}} G^{-1}(F([c]_\eta)^\circ)}.$$

Having only a finite number of sets $F([c]_\eta)^\circ$ we can, without loss of generality, assume that there is a sequence $w^k \to w$ with $w^k \in G^{-1}(F([c]_\eta)^\circ)$ for a single fixed set $F([c]_\eta)^\circ$. Then we have $S(w) = \lim_{k\to\infty} S(w^k) \in \partial_B H(w)$.

To complete the proof we assume that there is a $w \in \mathbb{R}^m \setminus R$. Then there is an open ball $B_\epsilon(w)$ such that $B_\epsilon(w) \cap G^{-1}(F([c]_\eta)^\circ) = \emptyset$ for all c, η. By the open mapping theorem (see, e.g., [108]) $G(B_\epsilon(w))$ is also open. Thus it must intersect at least one $F([c]_\eta)^\circ$ which contradicts the assumption and shows that $\mathbb{R}^m = R$. □

Remark 5.3. *While Proposition 5.5 seems to give a reasonable characterization of $S(w)$, the assumption* $\operatorname{rank} B = n$ *is quite restrictive for the following reason. If the saddle point problem arises from a minimization problem with linear constraints we have $C = 0$ in general, and a well posed problem will have $m \leq n$ linear constraints only. Combined with* $\operatorname{rank} B = n$ *this results in B to be a regular square matrix and hence $u = B^{-1}g$.*

The following example shows that the assertion of Proposition 5.5 is in general not valid if $\operatorname{rank} B < n$.

Example 5.2. *For $K = \{x \in \mathbb{R}^2 : x_i \geq 0, i = 1, 2\}$ consider the saddle point problem*

$$\begin{pmatrix} F & B^T \\ B & -C \end{pmatrix}\begin{pmatrix} u \\ w \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \partial\chi_K & \begin{pmatrix} -1 \\ 1 \end{pmatrix} \\ \begin{pmatrix} -1 & 1 \end{pmatrix} & -1 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ w \end{pmatrix} \ni \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}.$$

Then we have $(F^{-1}(y))_i = \max\{0, y_i\}$ and thus the nonlinear Schur complement is

$$H(w) = \max\{0, w\} - \max\{0, -w\} + w = 2w.$$

Hence $\nabla H(w) = 2$ and $\partial_C H(w) = \partial_B H(w) = \{2\}$. On the other hand we have $S(w) = 2$ for $w \neq 0$ but $S(0) = 1$.

This problem occurs since the line $f - B^T w$ crosses three domains where F^{-1} is smooth. While these domains have a nonempty interior themselves the line intersects the one leading to $F^{-1}(y) = 0$ only at the single point $y = 0$ of its border. Thus the preimage of this domain under $f - B^T(\cdot)$ collapses to the single point $w = 0$.

5.4 Schur Nonsmooth Newton Methods

We now consider the algorithms obtained if a linearization of the nonlinear Schur complement is used as preconditioner for search directions

$$d^\nu = -S(w^\nu)^{-1}\nabla h(w^\nu) \tag{5.42}$$

in the general descent algorithms given by (5.10), or (5.12) and (5.13). Convergence will follow from the convergence results for gradient-related descent methods.

5.4.1 Algorithms and Convergence

Before showing convergence of the algorithms we consider the solvability of the system

$$S(w^\nu)d^\nu = -\nabla h(w^\nu).$$

We immediately get

$$\langle S(w)x, y\rangle = \left\langle \left(\partial^2 J(u)_{\mathcal{I}'(u)}\right)^+ B^T x, B^T y \right\rangle + \langle Cx, y\rangle, \qquad x, y \in \mathbb{R}^m,$$

for $u = F^{-1}(f - B^T w)$. Hence by (A3), (A4), and (A10) the matrix $S(w)$ is symmetric and positive semidefinite. However we have no guarantee that it is invertible.

Even if this matrix is invertible it will often not be possible to solve the above system directly, and the application of iterative schemes does in general involve multiplications by $\left[\partial^2 J_0(u) + \varphi''(u)\right]^+_{\mathcal{I}'(u)}$. While this is in principle possible the possibly large derivatives φ''_i might prevent convergence. In order to overcome this problem recall the definition

$$\mathcal{I}''(v) := \{i \in \mathcal{I}'(v) : \varphi''_i(v_i) < (C_\varphi)_{i,i}\}$$

of the reduced inactive set introduced in (4.36) for a positive definite diagonal matrix $C_\varphi \in \mathbb{R}^{m,m}$. The induced truncated linearization of H at w with $u = F^{-1}(f - B^T w)$ is given by

$$S'(w) := B\Big(\partial^2 J(u)_{\mathcal{I}''(u)}\Big)^+ B^T + C.$$

This additional truncation of the matrix ensures that the diagonal elements of $\partial^2 J(u)_{\mathcal{I}''(u)}$ remain bounded independently of $\varphi''(u)$. Note that this involves the same truncated Hessian of J as already used for the TNNMG method in Corollary 4.3 and Corollary 4.4.

Independently of this truncation the matrix $S'(w)$ may not be invertible. In the most extreme case $S'(w) = 0$ if all components are active while the system results from a constraint minimization problem, i.e. $\mathcal{I}''(u) = \{1, \dots, n\}$ and $C = 0$. Although this does not happen in many application problems, it is not uncommon that $S'(w)$ has a nontrivial kernel.

Since the kernel of $\left[\partial^2 J_0(u) + \varphi''(u)\right]_{\mathcal{I}''(u)}$ and thus the kernel of $S'(w)$ with $u = F^{-1}(f - B^T w)$ depends only on $\mathcal{I}''(u)$, the same is true for the orthogonal projection $P_{\ker(S'(w))} : \mathbb{R}^m \to \ker(S'(w))$. Hence for a fixed symmetric positive definite matrix $\tilde{C}$ we can define the symmetric positive semidefinite matrix

$$\tilde{C}(\mathcal{I}''(u)) := P^T_{\ker(S'(w))} \tilde{C} P_{\ker(S'(w))} \in \mathbb{R}^{m,m},$$

and introduce the regularized linearization of H given by

$$S''(w) := S'(w) + \tilde{C}(\mathcal{I}''(u)).$$

If $v_{w,1}, \dots, v_{w,l}$ is an orthonormal basis of $\ker(S'(w))$ then it is easy to see that $P_{\ker(S'(w))}$ and $\tilde{C}(\mathcal{I}''(u))$ are given by

$$P_{\ker(S'(w))} = \sum_{i=1}^{l} v_{w,i} v_{w,i}^T, \qquad \tilde{C}(\mathcal{I}''(u)) = \sum_{i,j=1}^{l} \langle v_{w,i}, v_{w,j} \rangle_{\tilde{C}} \, v_{w,i} v_{w,j}^T.$$

Lemma 5.8. *$S''(w)$ is symmetric and positive definite for all $w \in \mathbb{R}^m$.*

Proof. Let $x_1, x_2 \in \mathbb{R}^m$ and $x_i^\bullet = P_{\ker(S'(w))} x_i$, $x_i^\circ = x_i - x_i^\bullet$. Then symmetry and definiteness follow from

$$\big\langle S''(w) x_1, x_2 \big\rangle = \big\langle S'(w) x_1^\circ, x_2^\circ \big\rangle + \Big\langle \tilde{C} x_1^\bullet, x_2^\bullet \Big\rangle.$$

□

Theorem 5.7. *The directions generated by* $d(\nu, w) = -S''(w)^{-1}\nabla h(w)$ *are gradient-related and guarantee* $\nabla h(v^\nu) \to 0$ *for any sequence* $v^\nu \in \mathbb{R}^m$ *with* $d(\nu, v^\nu) \to 0$.

Proof. The equivalence $d(\nu, w) = 0 \Leftrightarrow \nabla h(w) = 0$ in (5.16) follows from the fact that each $S''(w)$ is regular. To prove the estimate (5.17) let $w \in \mathbb{R}^m$ and define the reduced space

$$V_{\mathcal{I}} := \operatorname{span}\{e_i : i \in \mathcal{I}\} = \{v \in \mathbb{R}^n : v = v_{\mathcal{I}}\} \tag{5.43}$$

for any index set $\mathcal{I}$. For $u = F^{-1}(f - B^T w)$ we then have

$$\begin{aligned}\left\langle \partial^2 J(u)_{\mathcal{I}''(u)} v, v\right\rangle &\le \left\langle \overline{H}_{J_0} v, v\right\rangle + \langle C_\varphi v, v\rangle \\ &\le \lambda_{\max}(\overline{H}_{J_0} + C_\varphi)\, \langle v, v\rangle \qquad \forall v \in V_{\mathcal{I}''(u)},\end{aligned}$$

and

$$\begin{aligned}\left\langle \partial^2 J(u)_{\mathcal{I}''(u)} v, v\right\rangle &\ge \left\langle \underline{H}_{J_0} v, v\right\rangle \\ &\ge \lambda_{\min}(\underline{H}_{J_0})\, \langle v, v\rangle \qquad \forall v \in V_{\mathcal{I}''(u)}.\end{aligned}$$

Since the eigenvalues of $\partial^2 J(u) = \partial^2 J_0(u) + \varphi''(u)$ restricted to the indices in $\mathcal{I}''(u)$ are bounded, the same is true for the restricted inverse. Thus the following estimate holds for all $v \in \mathbb{R}^n$

$$\begin{aligned}\lambda_{\max}(\overline{H}_{J_0} + C_\varphi)^{-1} \left\langle I_{\mathcal{I}''(u)} v, v\right\rangle &\le \left\langle \left(\partial^2 J(u)_{\mathcal{I}''(u)}\right)^+ v, v\right\rangle \\ &\le \lambda_{\min}(\underline{H}_{J_0})^{-1} \left\langle I_{\mathcal{I}''(u)} v, v\right\rangle \\ &\le \lambda_{\min}(\underline{H}_{J_0})^{-1} \langle v, v\rangle .\end{aligned}$$

Using these estimates for $S''(w)$ we get for $v \in \mathbb{R}^m$

$$\begin{aligned}&\min\left\{\frac{1}{\lambda_{\max}(\overline{H}_{J_0} + C_\varphi)}, 1\right\} \left\langle \left(B I_{\mathcal{I}''(u)} B^T + C + \tilde{C}(\mathcal{I}''(u))\right) v, v\right\rangle \\ &\qquad \le \left\langle S''(w) v, v\right\rangle \le \max\left\{\frac{1}{\lambda_{\min}(\underline{H}_{J_0})}, 1\right\} \left\langle \left(B B^T + C + \tilde{C}(\mathcal{I}''(u))\right) v, v\right\rangle .\end{aligned}$$

Recalling that

$$\ker(I_{\mathcal{I}''(u)}) = \ker\left(\partial^2 J(u)_{\mathcal{I}''(u)}\right)^+$$

it is clear that the matrix on the left of the inequality is regular. Hence the matrices $S''(w)$ are bounded

$$\gamma_{\mathcal{I}''(u)} \langle v, v\rangle \le \left\langle S''(w) v, v\right\rangle \le \Gamma_{\mathcal{I}''(u)} \langle v, v\rangle$$

with constants $\gamma_{\mathcal{I}''(u)}, \Gamma_{\mathcal{I}''(u)} > 0$ depending only on the inactive set $\mathcal{I}''(u)$. Using this we get

$$\left\langle y, S''(w)^{-1} y\right\rangle \ge \gamma_{\mathcal{I}''(u)} \|S''(w)^{-1} y\|^2 \ge \frac{\gamma_{\mathcal{I}''(u)}}{\Gamma_{\mathcal{I}''(u)}} \|S''(w)^{-1} y\| \|v\|,$$

and thus (5.17) with

$$c_D = \min_{\mathcal{J}\subset\{1,\dots,n\}} \frac{\gamma_{\mathcal{J}}}{\Gamma_{\mathcal{J}}}.$$

Finally we note that we get $\nabla h(v^\nu) \to 0$ from

$$\|\nabla h(v^\nu)\| \leq \|S''(v^\nu)\| \, \|d(\nu, v^\nu)\| \leq \max_{\mathcal{J}\subset\{1,\dots,n\}} \Gamma_{\mathcal{J}} \, \|d(\nu, v^\nu)\|$$

for any sequence v^ν with $d(\nu, v^\nu) \to 0$. □

While this proof allows to apply the generic convergence results to the descent method obtained using the directions $d^\nu = -S''(w^\nu)^{-1}\nabla h(w^\nu)$ for the whole problem class, it is suboptimal in the following sense:

Since all estimates are derived for the Euclidean norm, the constant c_D incorporates the condition number of $\nabla^2 J_0(u)$, which may be large for discretized partial differential equations. For special cases it may be possible to derive much better estimates if a suitable norm for w is used. However, such improvements would only be visible in the convergence result of Theorem 5.3, since the more general result in Theorem 5.2 uses a compactness argument that does not give bounds.

Corollary 5.3. *Let $w^0 \in \mathbb{R}^m$. Then the sequence w^ν defined by*

$$\begin{aligned} d^\nu &= -S''(w^\nu)^{-1}\nabla h(w^\nu), \\ w^{\nu+\frac{1}{2}} &= w^\nu + \rho^\nu d^\nu, \\ w^{\nu+1} &= w^{\nu+\frac{1}{2}} + \mathcal{C}(w^{\nu+\frac{1}{2}}) \end{aligned}$$

converges to the solution w^* of (5.7) *if the step size rule $\rho^\nu = \rho(\nu, w^\nu, d^\nu)$ generates efficient step sizes.*

The same is true if d^ν is replaced by descent directions $\tilde{d}^\nu$ such that $\|d^\nu - \tilde{d}^\nu\|$ satisfies the accuracy condition (5.24) *of Proposition 5.2 and if ρ^ν is replaced by $\tilde{\rho}^\nu$ in the sense of Theorem 5.4.*

Proof. From Theorem 5.7 and Proposition 5.2 it follows that we have gradient-related descent directions. Thus we can apply Theorem 5.2 if $\rho^\nu = \rho(\nu, w^\nu, d^\nu)$ is chosen. If ρ^ν is replaced by $\tilde{\rho}^\nu$ in the sense of Theorem 5.4 we only have to note that $d(\nu, v^\nu) \to 0$ implies $\nabla h(v^\nu) \to 0$, by Theorem 5.7. □

The algorithm in Corollary 5.3 is essentially an inexact damped Newton-type method for the operator $H = \nabla h$. For $\mathcal{C} = 0$ it takes the form

$$w^{\nu+1} = w^\nu - \rho^\nu S''(w^\nu)^{-1} H(w^\nu)$$

with $\rho^\nu = \rho(\nu, w^\nu, -S''(w^\nu)^{-1}\nabla h(w^\nu))$. Since $S''(w)$ plays the role of a generalized linearization of the nonsmooth nonlinear Schur complement H at w we call this a "Schur Nonsmooth Newton method".

Lemma 5.9. *Let J_0 be twice continuously differentiable and let all φ_i'' be uniformly bounded from above by a constant $c_{\varphi''}$, whenever $\partial\varphi_i$ is single-valued. Then h is strongly convex if $S(w)$ is symmetric positive definite for all $w \in \mathbb{R}^m$.*

Proof. Let $w_1, w_2 \in \mathbb{R}^n$ and $x_i = f - B^T w_i$. By Theorem 5.6 we have for some $c > 0$

$$\begin{aligned}&\langle H(w_1) - H(w_2), w_1 - w_2\rangle\\ &\qquad = \left\langle F^{-1}(x_1) - F^{-1}(x_2), x_1 - x_2\right\rangle + \langle w_1 - w_2, w_1 - w_2\rangle_C\\ &\qquad \geq c\,\langle x_1 - x_2, x_1 - x_2\rangle_{\mathcal{I}_0} + \langle w_1 - w_2, w_1 - w_2\rangle_C\\ &\qquad = \langle w_1 - w_2, w_1 - w_2\rangle_{BcI_{\mathcal{I}_0}B^T + C}\,.\end{aligned}$$

Now let $x \in \operatorname{dom}\varphi$ such that $\mathcal{I}(x) = \mathcal{I}_0$ and $y \in F(x)$. Then the kernels of $(\partial^2 J(u)_{\mathcal{I}_0})^+$ and $cI_{\mathcal{I}_0}$ coincide. Thus the reduced Schur complement

$$BcI_{\mathcal{I}_0}B^T + C$$

must also be positive definite because $S(w)$ is. Hence $\nabla h = H$ is strongly monotone and h is strongly convex. □

Corollary 5.4. *Let J_0, φ, $S(w)$, and ρ satisfy the assumptions of Lemma 5.9 and Corollary 5.3, and let $(C_\varphi)_i > c_{\varphi''}$. Then $S(w) = S'(w) = S''(w)$ holds true and the method in Corollary 5.3 converges R-linearly.*

The same is true if d^ν is replaced by descent directions $\tilde{d}^\nu$ such that $\|d^\nu - \tilde{d}^\nu\|$ satisfies the accuracy condition (5.24) *of Proposition 5.2.*

Proof. Combine Theorem 5.7, Lemma 5.9 and Theorem 5.3. □

In general one would expect local superlinear convergence of a Newton-type method. Unfortunately our preconditioners $S''(w)$ are in general not contained in $\partial_C H(w)$ for the following reasons:

- As shown by Example 5.2 we may have $S'(w) \notin \partial_C H(w)$ if $\operatorname{rank} B \neq n$ due to the lack of a chain rule.
- If ∇J_0 is not differentiable it may not be possible to choose $\partial^2 J_0(w) \in \partial_C(\nabla J_0(w))$. Even if this is possible the lack of a chain rule may lead to $S'(w) \notin \partial_C H(w)$.
- In case of unbounded second derivatives of φ additional truncation is introduced.
- $S'(w)$ may not be invertible and thus needs to be regularized.

In all of the above cases the classical convergence analysis of semismooth Newton methods as introduced by Kummer [79], Pang [85], Qi and Sun [89] cannot be applied. The remaining case is considered in the following proposition.

Proposition 5.6. *Let J_0, φ, and $S(w)$ satisfy the assumptions of Lemma 5.9 and let $(C_\varphi)_i > c_{\varphi''}$ and $\operatorname{rank} B = n$. Then $S(w) = S'(w) = S''(w)$ holds true and the sequence w^ν defined by*

$$w^{\nu+1} = w^\nu - S''(w^\nu)^{-1}\nabla h(w^\nu) \tag{5.44}$$

converges superlinearly to the solution w^* of (5.7) *if $\|w^0 - w^*\|$ is small enough.*

Proof. By Proposition 5.5 we have

$$S''(w^\nu) = S(w^\nu) \in \partial_B H(w^\nu) \qquad \forall \nu \in \mathbb{N}.$$

Hence the method (5.44) is a classical nonsmooth Newton method as introduced in [89]. Since the second derivatives φ_i'' are bounded on $F([c]_\eta)^\circ$, each φ_i'' with $i \in \mathcal{I}(c)$ and thus F^{-1} can be extended continuously differentiable to a larger open set containing $\overline{F([c]_\eta)^\circ}$. This guarantees that F^{-1} is piecewise smooth [106, Definition 2.19] and semismooth [106, Proposition 2.26]. Thus the above method converges superlinearly in a sufficiently small neighborhood (see [89, Theorem 3.2], [106, Proposition 2.12]). □

This result is unsatisfactory not only because of the restrictive assumptions (cf. Remark 5.3). It also does not give any information on the domain of convergence.

Proposition 5.7. *Let the assumptions of Proposition 5.6 be satisfied and assume that the solution w^* of* (5.2) *satisfies the non-degeneracy condition*

$$\exists \eta^* \in \mathcal{E}: \qquad f - B^T w^* \in F([u^*, \eta^*])^\circ \tag{5.45}$$

with $u^* = F^{-1}(f - B^T w^*)$. Then (5.44) *reduces to a classical Newton method for H in the open neighborhood*

$$U := (f - B^T(\cdot))^{-1}(F([u^*, \eta^*])^\circ).$$

Analogously the method of Corollary 5.3 with $\mathcal{C} = 0$ reduces to a damped classical Newton method on U.

Proof. We only have to note that F^{-1} is differentiable on $F([u^*, \eta^*])^\circ$ and that $f - B^T(\cdot)$ is continuous. □

In view of Proposition 5.7 the result of Proposition 5.6 is almost useless. Provided that the non-degeneracy condition on w^* holds, one can simply apply the convergence theory for classical smooth Newton methods in a small neighborhood U' contained in U. Since Proposition 5.6 does not ensure that the domain of convergence is larger than U' it does not give any additional information. If the inactive set $\mathcal{I}(u^*)$ of w^* and the set $F([u^*, \eta^*])^\circ$ are not known, then there is no hope that the local result can be applied. Moreover the determination of $\mathcal{I}(u^*)$ and $F([u^*, \eta^*])^\circ$ is generally not a simpler task then solving the original problem.

Remark 5.4. *Although the convergence analysis only guarantees superlinear convergence of the undamped version in a neighborhood of unknown size or, under the nondegeneracy condition* (5.45)*, in a neighborhood where H is smooth, the method does in practice converge superlinearly on a much larger domain. While it is complicated to choose parameters for the Armijo rule such that $\rho^\nu \to 1$ and thus asymptotic coincidence with an undamped Newton method is guaranteed, this is in general the case for the bisection rule if ϵ is small enough.*

To ensure this behavior theoretically, nonsmooth analogues of well-known affine-invariant damping strategies (cf. Deuflhard [43]) are needed. The crucial part here is the development of a local convergence theory which is robust with respect to different regions of smoothness.

Remark 5.5. *For certain problems, where F^{-1} maps L^2 to L^2 there are convergence results of nonsmooth Newton methods in function spaces [64, 95, 96, 106]. Discrete analogues of these methods are asymptotically robust since the function space convergence theory does not rely on smooth regions.*

In contrast it is not clear how to show robustness in discrete cases that lack a function space analogue. The general problem is that continuous properties that hold on domains with asymptotically vanishing measure, translate into properties that do not hold for any discrete component asymptotically.

5.4.2 Computational Aspects

As already mentioned the terms h and $\nabla h = H$ are in general not explicitly available. In order to obtain an efficient method it is crucial to have fast iterative schemes to evaluate these quantities.

Before dealing with this problem we note that for $\mathcal{C} = 0$ the Schur Nonsmooth Newton method in Corollary 5.3 can equivalently written as

$$u^\nu = F^{-1}(f - B^T w^\nu), \tag{5.46}$$

$$w^{\nu+1} = w^\nu + \rho^\nu \underbrace{S''(w^\nu)^{-1}(Bu^\nu - Cw^\nu - g)}_{=:d^\nu} \tag{5.47}$$

with $\rho^\nu = \rho(\nu, w^\nu, d^\nu)$. This is a preconditioned Uzawa method for the original saddle point problem (5.1). If F is a linear operator it reduces to the classical Uzawa method and $S''(w)$ reduces to the linear Schur complement. In this case the preconditioned method obviously terminates within one step. If F is associated with a quadratic obstacle problem and the preconditioner is omitted standard convergence results for Uzawa methods can be applied yielding even an a priori fixed interval of allowed step sizes [55, 56].

The first substep amounts to the evaluation of F^{-1}, which is equivalent to the solution of the minimization problem

$$u^\nu = \operatorname*{arg\,min}_{u \in \mathbb{R}^n} \left(J(u) - \langle f - B^T w^\nu, u \rangle \right).$$

Since the assumptions made for the TNNMG method in Chapter 4 are a subset of the assumptions for the Schur Nonsmooth Newton method considered here this problem can be solved using the TNNMG method provided that a proper hierarchy of subspaces is given e.g. by a multigrid hierarchy in case of a discretized partial differential equation.

If the latter is not the case it is still possible to solve the reduced linear subproblems inexactly using other iterative methods like the preconditioned conjugate gradients method. While even the exact solution of the linear subproblems using a direct sparse solver (see, e.g., [39, 42]) is possible this will in general lead to an overall algorithm of suboptimal algorithmic complexity. Since the matrix is positive definite on an explicitly known subspace another alternative is to use an algebraic multigrid approach (see, e.g., [92]) to construct subspaces from the matrix only.

In the special case of a quadratic obstacle problem there are also various other methods [36, 54, 64] especially of multigrid or domain decomposition type [5, 6, 69, 80, 104]. For a comparison of the latter we refer to [58]. While some of those methods are restricted to the quadratic obstacle problem the methods in [5, 6, 54] can also be applied in case of a nonquadratic J_0 and the method in [69] has been extended to piecewise smooth φ [75].

The evaluation of F^{-1} is also needed if h or ∇h have to be evaluated in order to compute ρ^ν using a step size rule like the Armijo rule or bisection as discussed in Section 5.2.3. This leads to multiple evaluations of F^{-1} per iteration step in general. If this is expensive it may be advantageous to adaptively switch off the step rule using the criterion (5.33) of Theorem 5.4. In view of the interpretation of the method as a Newton-type method one can hope that the norms of the directions decrease for good initial iterates. In this case the step rule will not be switched on only one evaluation of F^{-1} remains. However, the adaptive criterion (5.33) ensures that the method does still converge globally if this is not the case.

The second substep (5.47) involves the evaluation of $S''(w^\nu)^{-1}$. It can be written as the linear saddle point problem

$$\overline{u}^\nu \in \mathbb{R}^n,\ d^\nu \in \mathbb{R}^m : \qquad \begin{pmatrix} A^\nu & (B^\nu)^T \\ B^\nu & -C^\nu \end{pmatrix} \begin{pmatrix} \overline{u}^\nu \\ d^\nu \end{pmatrix} = \begin{pmatrix} 0 \\ g^\nu \end{pmatrix} \tag{5.48}$$

with

$$\begin{aligned} A^\nu &= \Big(\partial^2 J_0(u) + \varphi''(u)\Big)_{\mathcal{I}''(u^\nu)}, \\ B^\nu &= B_{\mathbb{N},\mathcal{I}''(u^\nu)}, \\ C^\nu &= C + \tilde{C}(\mathcal{I}''(u^\nu)), \\ g^\nu &= \nabla h(w^\nu) = g + Cw^\nu - Bu^\nu, \end{aligned}$$

for an auxiliary variable $\overline{u}^\nu$. Since A^ν represents a linearization of $F = \partial J$ on the reduced space

$$V_{\mathcal{I}''(u^\nu)} = \operatorname{span}\{e_i : i \in \mathcal{I}''(u^\nu)\} = \mathbb{R}^n / \ker A^\nu, \tag{5.49}$$

this system can be regarded as a regularized linearization of the saddle point problem (5.1) on the reduced space $V_{\mathcal{I}''(u^\nu)} \times \mathbb{R}^m$. By construction the linear Schur complement of (5.48) is given by

$$S''(w^\nu) = B(A^\nu)^+ B^T + C + \tilde{C}(\mathcal{I}''(u^\nu)) = B^\nu (A^\nu)^+ (B^\nu)^T + C + \tilde{C}(\mathcal{I}''(u^\nu)).$$

Proposition 5.8. *The linear saddle point problem* (5.48) *has a unique solution* $(\overline{u}^\nu, d^\nu) \in V_{\mathcal{I}''(u^\nu)} \times \mathbb{R}^m$ *given by* $d^\nu = -S''(w^\nu)^{-1} g^\nu$ *and* $\overline{u}^\nu = -(A^\nu)^+ (B^\nu)^T d^\nu$. *The solutions of* (5.48) *in* $\mathbb{R}^n \times \mathbb{R}^m$ *are given by* $(\overline{u}^\nu + v^\nu, d^\nu) \in \mathbb{R}^n \times \mathbb{R}^m$ *with* $v^\nu \in V^\perp_{\mathcal{I}''(u^\nu)} = \{v \in \mathbb{R}^n : v = v_{\mathcal{I}''(u^\nu)}\} = \ker A^\nu$.

Proof. Replace A^ν by $A^\nu + I - I_{\mathcal{I}''(u^\nu)}$ and the right hand side of the first equation by v with $v \in V^\perp_{\mathcal{I}''(u^\nu)}$. Then a simple block elimination yields that $(\overline{u}^\nu + v, d^\nu) \in \mathbb{R}^n \times \mathbb{R}^m$ with $(\overline{u}^\nu, d^\nu)$ as given above is the unique solution of this modified system. Now Lemma A.5 together with the invariance of the original system under modifications $\overline{u}^\nu + v$ with $v \in \ker A^\nu$ provide the assertion. □

In view of this result the solution of (5.48) can either be obtained by considering the system on the subspace $V_{\mathcal{I}''(u^\nu)} \times \mathbb{R}^m = (\mathbb{R}^n / \ker A^\nu) \times \mathbb{R}^m$ only or by adding the orthogonal projection onto the kernel given by $P_{\ker A^\nu} = I - I_{\mathcal{I}''(u^\nu)} = I_{\mathbb{N} \setminus \mathcal{I}''(u^\nu)}$ to A^ν in order to make the part of $\overline{u}^\nu$ in $V^\perp_{\mathcal{I}''(u^\nu)}$ unique.

While there are general methods to solve the nonlinear convex minimization problems associated with F^{-1} the situation looks different for the linear saddle point problem. Since the problem is linear and symmetric it is possible to use a direct solver or Krylov methods like GMRES [93] or MINRES [84]. Due to the indefinite matrix there is no general multigrid method. However, there are multigrid methods that work well in special cases. Some of those methods require the saddle point problem to be related to a quadratic minimization problem with linear constraints, i.e. $C^\nu = 0$. Since this does not hold in general for the subproblems (5.48) we note that they can also be reformulated in the following way.

Proposition 5.9. *The linear saddle point problem* (5.48) *is equivalent to the saddle point problem*

$$\overline{u}^\nu \in \mathbb{R}^n, d_0^\nu \in \mathbb{R}^m, d^\nu \in \mathbb{R}^m : \qquad \begin{pmatrix} A^\nu & 0 & (B^\nu)^T \\ 0 & C^\nu & -C^\nu \\ B^\nu & -C^\nu & 0 \end{pmatrix} \begin{pmatrix} \overline{u}^\nu \\ d_0^\nu \\ d^\nu \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ g^\nu \end{pmatrix} \tag{5.50}$$

in the sense that $(\overline{u}^\nu, d_0^\nu, d^\nu)$ *is a solution of* (5.50) *iff* $(\overline{u}^\nu, d^\nu)$ *is a solution of* (5.48) *and* $Cd_0^\nu = Cd^\nu$. *The solutions of* (5.50) *are unique in* $V_{\mathcal{I}''(u^\nu)} \times (\mathbb{R}^m / \ker C) \times \mathbb{R}^m$ *and the Schur complement is given by* $S''(w^\nu)$.

Again we can construct a system that is uniquely solvable in $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$ by adding $P_{\ker A^\nu}$ to A^ν and $P_{\ker C^\nu}$ to the appearance of C^ν on the diagonal of (5.50) without changing the part of the solution in $V_{\mathcal{I}''(u^\nu)} \times (\mathbb{R}^m / \ker C) \times \mathbb{R}^m$.

One class of multigrid methods for systems of the form (5.50) uses the smoother by Braess and Sarazin [21]. Each application of this smoother incorporates the solution of a linear problem for $(B^\nu - C^\nu)((B^\nu)^T - C^\nu)^T$. While this reduces to a discretized second order elliptic problem for the Stokes problem it is not appropriate if B^ν or C^ν themselves result from a second order differential operator.

Another approach is to construct a smoother by successively solutving small local saddle point problems that couple only a few primal and dual unknowns in a so-called patch. Such smoothers were introduced by Vanka [107] for the Navier–Stokes equations. For the case of a parallel solution of the local problems, i.e. block Jacobi patch smoothers, convergence results were established by Zulehner [116, 117], Schöberl and Zulehner [98], and Simon and Zulehner [101].

Since the Schur Nonsmooth Newton method is robust with respect to inexact evaluation of $d^\nu = -S''(w^\nu)^{-1}\nabla h(w^\nu)$ it is in general only necessary to solve the linear saddle point problem (5.48) inexactly. Furthermore, the method does still converge if $S''(w^\nu)$ is replaced by $S''(\tilde{w}^\nu)$ for some approximation $\tilde{w}^\nu$ of w^ν, due to the uniform boundedness of $S''(w)$ with respect to w. Noting that $S''(w)$ does, in fact, only depend on $u = F^{-1}(f - G^T w)$ this allows to replace $S''(w^\nu)$ by

$$S''(\tilde{w}^\nu) = B\Big(\partial^2 J_0(\tilde{u}^\nu) + \varphi''(\tilde{u}^\nu)\Big)^+_{\mathcal{I}''(\tilde{u}^\nu)} B^T + C + \tilde{C}(\mathcal{I}''(\tilde{u}^\nu))$$

for any approximation $\tilde{u}^\nu$ of u^ν. Defining

$$d^\nu = -S''(\tilde{w}^\nu)^{-1} \underbrace{(g + Cw^\nu - Bu^\nu)}_{=\nabla h(w^\nu)}, \qquad \tilde{d}^\nu = -S''(\tilde{w}^\nu)^{-1} \underbrace{(g + Cw^\nu - B\tilde{u}^\nu)}_{\widetilde{\nabla h(w^\nu)}} + \epsilon^\nu$$

we know from Proposition 5.2 that the Schur Nonsmooth Newton method of Corollory 5.3 converges as long as

$$\frac{\|d^\nu - \tilde{d}^\nu\|_M}{\|\tilde{d}^\nu\|_M} \leq \|S''(\tilde{w}^\nu)^{-1}\| \frac{\|B(u^\nu - \tilde{u}^\nu)\|_{M^{-1}}}{\|\tilde{d}^\nu\|_M} + \frac{\|\epsilon^\nu\|_M}{\|\tilde{d}^\nu\|_M} \to 0.$$

While convergence of the second term can be monitored if an estimator for the algebraic error during the linear saddle point solver is available convergence of the first term cannot be guaranteed a priori since $\tilde{d}^\nu$ is computed after $\tilde{u}^\nu$. However, it can be checked a posteriori. If the tolerance for the computation of $\tilde{u}^\nu \approx u^\nu$ is chosen reasonably a recomputation of a better $\tilde{u}^\nu$ will rarely be necessary. It is even possible to directly check whether

$$-\Big\langle \widetilde{\nabla h(w^\nu)}, \tilde{d}^\nu \Big\rangle \geq \tilde{c}_D \|\widetilde{\nabla h(w^\nu)}\|_{M^{-1}} \|\tilde{d}^\nu\|_M$$

holds for some fixed guess $\tilde{c}_D$ as long as the following terms tend to zero

$$\begin{aligned} &\left\| \frac{\nabla h(w^\nu)}{\|\nabla h(w^\nu)\|_{M^{-1}}} - \frac{\widetilde{\nabla h(w^\nu)}}{\|\widetilde{\nabla h(w^\nu)}\|_{M^{-1}}} \right\|_{M^{-1}} \\ &\qquad \leq 2 \frac{\|\nabla h(w^\nu) - \widetilde{\nabla h(w^\nu)}\|_{M^{-1}}}{\|\widetilde{\nabla h(w^\nu)}\|_{M^{-1}}} = 2 \frac{\|B(u^\nu - \tilde{u}^\nu)\|_{M^{-1}}}{\|g + Cw^\nu - B\tilde{u}^\nu\|_{M^{-1}}} \to 0. \end{aligned}$$

In the special case that the energy J is related to a quadratic obstacle problem as in Example 4.1 the matrices in (5.48) only depend on the inactive set $\mathcal{I}''(u^\nu) = \mathcal{I}'(u^\nu) = \mathcal{I}(u^\nu)$. The only dependency on $u^\nu = F^{-1}(f - B^T w^\nu)$ appears in the right hand side of this system. Due to the special structure of F^{-1} it can be eliminated in the following way.

Proposition 5.10. *Let J_0 satisfy (A6) and φ be given as in Example 4.1. Then the unique solution $(\tilde{u}^\nu, \tilde{w}^\nu) \in V_{\mathcal{I}''(u^\nu)} \times \mathbb{R}^m$ of the linear saddle point problem*

$$\begin{pmatrix} A^\nu & (B^\nu)^T \\ B^\nu & -C^\nu \end{pmatrix} \begin{pmatrix} \tilde{u}^\nu \\ \tilde{w}^\nu \end{pmatrix} = \begin{pmatrix} (f - b - Au^\nu_{\mathcal{A}(u^\nu)})_{\mathcal{I}(u^\nu)} \\ g - Bu^\nu_{\mathcal{A}(u^\nu)} - \tilde{C}(\mathcal{I}(u^\nu))w^\nu \end{pmatrix} \tag{5.51}$$

satisfies

$$d^\nu = -S''(w^\nu)^{-1} \nabla h(w^\nu) = \tilde{w}^\nu - w^n.$$

Proof. First we note that the solution is unique by the arguments in Proposition 5.8. From the variational inequality (4.9) we instantly get that $u^\nu = F^{-1}(f - B^T w^\nu)$ satisfies

$$A^\nu u^\nu + (Au^\nu_{\mathcal{A}(u^\nu)})_{\mathcal{I}(u^\nu)} = (Au^\nu)_{\mathcal{I}(u^\nu)} = (f + b - B^T w^\nu)_{\mathcal{I}(u^\nu)}.$$

Hence $u^\nu_{\mathcal{I}(u^\nu)}$ is given by

$$u^\nu_{\mathcal{I}(u^\nu)} = (A^\nu)^+ \left(f + b - B^T w^\nu - Au^\nu_{\mathcal{A}(u^\nu)} \right).$$

Now let $d^\nu = -S''(w^\nu)^{-1} \nabla h(w^\nu)$ and $\tilde{w}^\nu = w^\nu + d^\nu$. Then

$$\begin{aligned} S''(w^\nu)\tilde{w}^\nu &= (B(A^\nu)^+ B^T + C + \tilde{C}(\mathcal{I}(u^\nu)))w^\nu - \nabla h(w^\nu) \\ &= B(A^\nu)^+ B^T w^\nu + Bu^\nu_{\mathcal{I}(u^\nu)} - \left(g - Bu^\nu_{\mathcal{A}(u^\nu)} - \tilde{C}(\mathcal{I}(u^\nu))w^\nu \right) \\ &= B(A^\nu)^+ \left(f + b - Au^\nu_{\mathcal{A}(u^\nu)} \right)_{\mathcal{I}(u^\nu)} - \left(g - Bu^\nu_{\mathcal{A}(u^\nu)} - \tilde{C}(\mathcal{I}(u^\nu))w^\nu \right). \end{aligned}$$

Thus $\tilde{w}^\nu = w^\nu + d^\nu$ is the second component of the solution of (5.51). □

By Proposition 5.10 it is sufficient to know the inactive set $\mathcal{I}(u^\nu)$ and the values of u^ν in the active components $\mathcal{A}(u^\nu)$ in order to compute the direction d^ν. This is because the obstacle problem associated with F^{-1} reduces to a linear problem for the inactive components. This linear problem for $u^\nu_{\mathcal{I}(u^\nu)}$ is incorporated in the linear saddle point problem by a modified right hand side while the structure of the linear saddle point problem given by the matrices remains the same. Since the computation of the active components is in general cheaper than the solution of the whole minimization problem this allows to improve the performance of the algorithm. If the problem is non-degenerate in the sense that

$$\exists \eta^\nu \in \mathcal{E} : \qquad f - B^T w^\nu \in F([u^\nu, \eta^\nu])^\circ \tag{5.52}$$

many iterative methods like, e.g., the Gauß–Seidel method even detect the active components after a finite number of steps [72].

5.4.3 Relation to Primal–Dual Active Set Methods

The primal–dual active set method introduced in Section 4.2.6 was first stated for quadratic obstacle problems without linear constraints. However, it is often also applied to problems with additional linear constraints $Bu = g$ having the special structure that B decomposes according to

$$\begin{pmatrix} B_1 & B_2 \end{pmatrix} = B,$$

with a regular matrix $B_2 \in \mathbb{R}^{m,m}$. This structure implies that the linear constraint can be eliminated by explicit restriction to the subspace where $Bu = g$ holds.

Besides this decomposition we will assume in this subsection that $C = 0$ holds and that J results from a quadratic obstacle problem as in Example 4.1 without lower obstacle, i.e. $\underline{\psi} = -\infty$, and without any obstacle in the i-th components with $i > n_1 = n - m > 0$. If we also split u^*, f, A and the convex set

$$K = K_1 \times \mathbb{R}^m = \{u_1 \in \mathbb{R}^{n_1} : u_1 \leq \overline{\psi_1}\} \times \mathbb{R}^m$$

according to the splitting of B the saddle point problem (5.1) takes the form

$$\begin{array}{lcl} u_1^* & \in & \mathbb{R}^{n_1} \\ u_2^* & \in & \mathbb{R}^m \\ w* & \in & \mathbb{R}^m \end{array} : \begin{pmatrix} A_{11} + \partial\chi_{K_1} & A_{12} & B_1^T \\ A_{21} & A_{22} & B_2^T \\ B_1 & B_2 & 0 \end{pmatrix} \begin{pmatrix} u_1^* \\ u_2^* \\ w^* \end{pmatrix} \ni \begin{pmatrix} f_1 \\ f_2 \\ g \end{pmatrix}. \tag{5.53}$$

Under the above assumptions we then have $u_2^* = B_2^{-1}(g - B_1 u_1^*)$.

Lemma 5.10. *The saddle point problem* (5.53) *is equivalent to the box-constrained quadratic minimization problem*

$$u_1 \in \mathbb{R}^{n_1} : \qquad \tilde{J}(u_1) \leq \tilde{J}(v) \qquad \forall v \in \mathbb{R}^{n_1} \tag{5.54}$$

where $\tilde{J}$ is given by

$$\tilde{J}(v) = J(v, B_2^{-1}(g - B_1 v)) = \frac{1}{2}\langle Mv, v\rangle - \langle b, v\rangle + \chi_{K_1}(v),$$

with the symmetric positive definite matrix

$$M = A_{11} - (A_{12}B_2^{-1}B_1) - (A_{12}B_2^{-1}B_1)^T + (B_2^{-1}B_1)^T A_{22}(B_2^{-1}B_1)$$

and the right hand side

$$b = f_1 - (B_2^{-1}B_1)^T(f_2 - A_{22}B_2^{-1}g) - A_{12}B_2^{-1}g.$$

Proof. Since (5.53) is equivalent to the minimization of J on the affine subspace where $Bu = g$ holds the equivalence follows directly from the fact that $\tilde{J}$ represents the restriction of J to this subspace. The representations of M and b follow from elementary computations. Symmetry and definiteness of M follow from symmetry and definiteness of A. □

Proposition 5.11. *The primal–dual active set method for the minimization problem* (5.54) *is equivalent to*

$$\begin{pmatrix} A_{11} & A_{12} & B_1^T \\ A_{21} & A_{22} & B_2^T \\ B_1 & B_2 & 0 \end{pmatrix} \begin{pmatrix} u_1^{\nu+1} \\ u_2^{\nu+1} \\ w^{\nu+1} \end{pmatrix} = \begin{pmatrix} f_1 - \lambda^{\nu+1} \\ f_2 \\ g \end{pmatrix}, \tag{5.55}$$

$$(u_1^{\nu+1})_{\mathcal{A}_\nu} = (\overline{\psi}_1)_{\mathcal{A}_\nu}, \tag{5.56}$$

$$(\lambda^{\nu+1})_{\mathcal{I}_\nu} = 0, \tag{5.57}$$

with the active set $\mathcal{A}_\nu = \{i : (\lambda^\nu + c(u_1^\nu - \overline{\psi}_1))_i > 0\}$ *and the inactive set* $\mathcal{I}_\nu = \mathbb{N} \setminus \mathcal{A}_\nu$.

Proof. Let (u_1^ν, λ^ν) be the iterates produced by the primal–dual active set method (4.44) and (4.45) for the minimization problem (5.54). Define the variables

$$\begin{aligned} u_2^\nu &:= B_2^{-1}(g - B_1 u_1^\nu), \\ w^\nu &:= B_2^{-T}(f_2 - A_{21}u_1^\nu - A_{22}u_2^\nu). \end{aligned}$$

Then the second and third equation in (5.55) hold and from (4.45) we get

$$\begin{aligned} -\lambda^{\nu+1} &= Mu_1^{\nu+1} - b \\ &= A_{11}u_1^{\nu+1} - f_1 + A_{12}B_2^{-1}(g - B_1u_1^{\nu+1}) \\ &\qquad + B_1^T B_2^{-T}\left[f_2 - A_{21}u_1^{\nu+1} - A_{22}B_2^{-1}(g - B_1u_1^{\nu+1})\right] \\ &= A_{11}u_1^{\nu+1} + A_{12}u_2^{\nu+1} + B_1^T w^{\nu+1} - f_1. \end{aligned}$$

Thus (5.55) is equivalent to (4.45). □

The system (5.55) still couples $u^{\nu+1}$ and $\lambda^{\nu+1}$. Due to its special structure we can eliminate $\lambda^{\nu+1}$.

Proposition 5.12. *The primal–dual active set method for the minimization problem* (5.54) *is equivalent to*

$$\begin{pmatrix} A_{\mathcal{I}_\nu} & (B_{\mathbb{N},\mathcal{I}_\nu})^T \\ B_{\mathbb{N},\mathcal{I}_\nu} & 0 \end{pmatrix} \begin{pmatrix} u^{\nu+1} \\ w^{\nu+1} \end{pmatrix} = \begin{pmatrix} (f - A\overline{\psi}_{\mathcal{A}_\nu})_{\mathcal{I}_\nu} \\ g - B\overline{\psi}_{\mathcal{A}_\nu} \end{pmatrix}, \tag{5.58}$$

$$u_{\mathcal{A}_\nu}^{\nu+1} = \overline{\psi}_{\mathcal{A}_\nu} \tag{5.59}$$

with $u^\nu = (u_1^\nu, u_2^\nu)$*, the active sets* $\mathcal{A}_0 = \{i : (\lambda^0 + c(u_1^0 - \overline{\psi}_1))_i > 0\}$ *and*

$$\mathcal{A}_\nu = \{i : (f - Au^\nu - B^T w^\nu + c(u^\nu - \overline{\psi}))_i > 0\},$$

and the inactive sets $\mathcal{I}_\nu = \mathbb{N} \setminus \mathcal{A}_\nu$.

Proof. We start by assuming that the iterates are given by the primal–dual active set method in the form of Proposition 5.11. Using $u_{\mathcal{A}_\nu}^{\nu+1} = \overline{\psi}_{\mathcal{A}_\nu}$ we get the splitting

$$(Au^{\nu+1})_{\mathcal{I}_\nu} = A_{\mathcal{I}_\nu}u^{\nu+1} + A_{\mathcal{I}_\nu,\mathcal{A}_\nu}\overline{\psi} = A_{\mathcal{I}_\nu}u^{\nu+1} + (A\overline{\psi}_{\mathcal{A}_\nu})_{\mathcal{I}_\nu}.$$

Thus the first equation in (5.58) is just the restriction of the first two equations in (5.55) to the indices in $\mathcal{I}_\nu$. The splitting

$$Bu^\nu = B_{\mathbb{N},\mathcal{I}_\nu} u^{\nu+1} + B_{\mathbb{N},\mathcal{A}_\nu} \overline{\psi}_{\mathcal{A}_\nu} = B_{\mathbb{N},\mathcal{I}_\nu} u^{\nu+1} + B\overline{\psi}_{\mathcal{A}_\nu}$$

implies the second equation. For the opposite direction we only need to note that the first equation in (5.58) implies (5.57) with

$$\lambda^\nu = f_1 - A_{11} u_1^\nu - A_{12} u_2^\nu - B_1^T w^\nu.$$

□

Theorem 5.8. *The primal–dual active set method for the minimization problem* (5.54) *is equivalent to*

$$w^{\nu+1} = w^\nu - (BA_{\mathcal{I}_\nu}^+ B^T)^{-1}(g - B\overline{u}^\nu) \tag{5.60}$$

with arbitrary $w^0 \in \mathbb{R}^m$, $\overline{u}^\nu$ *given by*

$$\overline{u}_{\mathcal{A}_\nu}^\nu = \overline{\psi}_{\mathcal{A}_\nu}, \qquad \overline{u}_{\mathcal{I}_\nu}^\nu = A_{\mathcal{I}_\nu}^+ (f - A\overline{\psi}_{\mathcal{A}_\nu} - B^T w^\nu)$$

and $\mathcal{A}_\nu$ *and* $\mathcal{I}_\nu$ *as defined in Proposition 5.12.*

Proof. By definition of $\overline{u}^\nu$ we have

$$\begin{aligned} A_{\mathcal{I}_\nu} \overline{u}_{\mathcal{I}_\nu}^\nu &= (f - A\overline{\psi}_{\mathcal{A}_\nu} - B^T w^\nu)_{\mathcal{I}_\nu}, \\ B\overline{\psi}_{\mathcal{A}_\nu} + B_{\mathbb{N},\mathcal{I}_\nu} \overline{u}^\nu &= B\overline{\psi}_{\mathcal{A}_\nu} + B\overline{u}_{\mathcal{I}_\nu}^\nu = B\overline{u}^\nu. \end{aligned}$$

Using this and Proposition 5.12 we get

$$\begin{pmatrix} A_{\mathcal{I}_\nu} & (B_{\mathbb{N},\mathcal{I}_\nu})^T \\ B_{\mathbb{N},\mathcal{I}_\nu} & 0 \end{pmatrix} \begin{pmatrix} u^{\nu+1} - \overline{u}^\nu \\ w^{\nu+1} - w^\nu \end{pmatrix} = \begin{pmatrix} 0 \\ g - B\overline{u}^\nu \end{pmatrix}.$$

Now (5.60) follows from block elimination. □

This formulation of the primal–dual active set method looks very similar to the Schur Nonsmooth Newton method. The question is if the sets of inactive indices of both methods coincide and if $\overline{u}^\nu = F^{-1}(f - B^T w^\nu)$ is true. In fact if $\mathcal{I}_\nu = \mathcal{I}(F^{-1}(f - B^T w^\nu))$ is true then $\overline{u}^\nu$ takes the same values as $F^{-1}(f - B^T w^\nu)$ for the active indices in $\mathcal{A}_\nu$ and solves the same linear equation for the inactive indices in $\mathcal{I}_\nu$ as $F^{-1}(f - B^T w^\nu)$. Hence in this case the method reduces to

$$w^{\nu+1} = w^\nu - (BA_{\mathcal{I}_\nu}^+ B^T)^{-1} \nabla h(w^\nu)$$

which is the undamped version of the Schur Nonsmooth Newton method. However, it is in general not true that the inactive sets coincide. An important special case where the active sets almost do coincide is given in the following theorem.

Theorem 5.9. *Assume that A is diagonal in the first n_1 lines and columns, i.e. A_{11} is diagonal and $A_{12} = 0 = A_{21}$. Then for $\nu \geq 1$ the primal–dual active set method for the minimization problem (5.54) is equivalent to*

$$w^{\nu+1} = w^\nu - (BA^+_{\mathcal{I}_\nu} B^T)^{-1} \nabla h(w^\nu)$$

for an arbitrary $w^0 \in \mathbb{R}^m$ and the inactive set

$$\mathcal{I}_\nu = \mathcal{I}(F^{-1}(f - B^T w^\nu)) \cup \{i : (F^{-1}(f - B^T w^\nu))_i = A^{-1}_{ii}(f - B^T w^\nu)\}.$$

Proof. Let $\nu \geq 1$ and $i \in \mathcal{A}_\nu \subset \{1, \ldots, n_1\}$. Then either $i \in \mathcal{A}_\nu \cap \mathcal{A}_{\nu-1}$ or $i \in \mathcal{A}_\nu \setminus \mathcal{A}_{\nu-1}$ and hence one of the following conditions

$$(u^\nu - \overline{\psi})_i = 0, \qquad \text{or} \qquad (f - B^T w^\nu)_i - A_{ii} u^\nu_i = 0, \tag{5.61}$$

is true, where A_{ii} now denotes the i-th diagonal entry of A. For the second condition we used the representation

$$\lambda^\nu_i = (f - Au^\nu - B^T w^\nu)_i = (f - B^T w^\nu)_i - A_{ii} u^\nu_i$$

of the active components of the residual λ^ν which follows directly from the assumptions on A. By (5.61) we instantly get

$$\begin{aligned} i \in \mathcal{A}_\nu &\Leftrightarrow i \in \mathcal{A}_\nu \cap \mathcal{A}_{\nu-1} \text{ or } i \in \mathcal{A}_\nu \setminus \mathcal{A}_{\nu-1} \\ &\Leftrightarrow \lambda^\nu_i > 0 \text{ or } (u^\nu - \overline{\psi})_i > 0 \\ &\Leftrightarrow A^{-1}_{ii} \lambda^\nu_i > 0 \text{ or } (u^\nu - \overline{\psi})_i > 0. \end{aligned}$$

Again by (5.61) $\lambda^\nu_i > 0$ is equivalent to

$$A^{-1}_{ii}(f - B^T w^\nu)_i - u^\nu_i = A^{-1}_{ii} \lambda^\nu_i > 0 = (\overline{\psi} - u^\nu)_i$$

and $(u^\nu - \overline{\psi})_i > 0$ is equivalent to

$$A^{-1}_{ii}(f - B^T w^\nu)_i - u^\nu_i = 0 > (\overline{\psi} - u^\nu)_i.$$

Hence $i \in \mathcal{A}_\nu$ is equivalent to

$$A^{-1}_{ii}(f - B^T w^\nu)_i > \overline{\psi}_i$$

where either the right hand side or the left-hand side is equals u^ν_i. Thus the inactive set is given by

$$\mathcal{I}_\nu = \mathcal{I}(F^{-1}(f - B^T w^\nu)) \cup \{i : (F^{-1}(f - B^T w^\nu))_i = A^{-1}_{ii}(f - B^T w^\nu)\}.$$

Now it is clear that $F^{-1}(f - B^T w^\nu)$ satisfies

$$\begin{aligned} (F^{-1}(f - B^T w^\nu))_{\mathcal{A}_\nu} &= \overline{\psi}_{\mathcal{A}_\nu}, \\ (F^{-1}(f - B^T w^\nu))_{\mathcal{I}_\nu} &= A^+_{\mathcal{I}_\nu}(f - B^T w^\nu) \end{aligned}$$

which implies $\overline{u}^\nu = F^{-1}(f - B^T w^\nu)$ and $g - B\overline{u}^\nu = \nabla h(w^\nu)$. □

In the case covered by Theorem 5.9 the primal–dual active set method almost coincides with the Schur Newton method. The only difference is the enlarged inactive set which also incorporates the indices i satisfying

$$(F^{-1}(f - B^T w^\nu))_i = A_{ii}^{-1}(f - B^T w^\nu).$$

These are the indices where strict complementarity does not hold, i.e., the solution of the local unconstrained problems coincides with the obstacle. This set will in general be small. If the selection of the active set in the primal–dual active set method is modified according to

$$\widetilde{\mathcal{A}}_\nu = \{i : \lambda_i^\nu + c(u_i^\nu - \overline{\psi}_i) \geq 0\},$$

as already done in Theorem 4.5, this difference also disappears. Hence, if A and B have the desired special structure, the Schur Nonsmooth Newton method can be regarded as a natural globalization of the primal–dual active set method. This was first discovered in [57] for the special case of an optimal control problem for a partial differential equation with control constraints and an L^2 regularization for the control.

6 Adaptive Numerical Solution of Nonlinear Saddle Point Problems

We have already introduced spatial discretizations for the time discrete Cahn–Hilliard equation on a given grid in Chapter 3 and algebraic solvers for the efficient solution of the obtained algebraic saddle point problems in Chapter 5. This chapter is dedicated to the adaptive numerical solution of the time discrete problems.

First we describe a local error indicator and a refinement strategy used to construct locally adaptive grids for the solution of nonlinear saddle point problems. Then we discuss implementation techniques and data structures needed for the spatial problems resulting from time discretization of an evolution problem.

6.1 Hierarchical Error Estimation for Nonlinear Saddle Point Problems

In Section 3.4 we introduced Rothe's method for the discretization of the Cahn–Hilliard equation from Chapter 2. The approach allows for time-dependent adaptive grids. This section will deal with the construction of these grids using hierarchical a posteriori error estimators.

Hierarchical error estimators for finite element discretizations were first introduced by Deuflhard et al. [45] and Zienkiewicz et al. [114]. The main idea is to extend the ansatz space and to estimate the error using the difference of the current approximation to an approximation in the extended space. For linear problems proper hierarchical preconditioning allows to compute an improved approximation in the extended space by local defect problems. The resulting local contributions can then be used as error indicators for adaptive grid refinement. For an overview we refer to the monograph of Ainsworth and Oden [1]. For linear symmetric problems local lower bounds can typically be shown without any unknown constants while upper bounds were established using local equivalence to residual estimators by [47] (see also [19]).

Due to their robustness and simplicity hierarchical error estimators are interesting for nonsmooth nonlinear problems. They have successively been applied to elliptic obstacle problems [66, 78, 100, 115] and nonsmooth convex minimization problems [61, 70, 94]. The application to quadratic saddle point problems with a superposition operator originating from the Cahn–Hilliard equation has been discussed in [61]. In the following we will present an extension that can be applied to time discrete anisotropic Cahn–Hilliard equations with logarithmic or obstacle potential.

Let $J_0 : H^1(\Omega) \to \mathbb{R}$ be a strongly convex differentiable functional with $\mathcal{F} = \nabla J_0$, and $\psi : H^1(\Omega) \to \mathbb{R} \cup \{\infty\}$ the convex, proper, and lower semicontinuous functional given by

$$\psi(v) := \int_\Omega \Psi(v(x))\, dx, \qquad\qquad \Psi : \mathbb{R} \to \mathbb{R} \cup \{\infty\},$$

for a convex, proper, and lower semicontinuous functional Ψ. Furthermore, let $b : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ be a symmetric bilinear form, $c : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ a symmetric positive semidefinite bilinear form, and $f, g \in L^2(\Omega)$ given right hand side functions. For these quantities consider the general saddle point problem:

Problem 6.1. *Find $(u, w) \in H^1(\Omega) \times H^1(\Omega)$ such that*

$$\begin{aligned} \langle \mathcal{F}(u), v - u\rangle + \psi(v) - \psi(u) + b(v - u, w) &\geq (f, v - u) \qquad && \forall v \in H^1(\Omega),\\ b(u, v) - c(w, v) &= (g, v) && \forall v \in H^1(\Omega). \end{aligned}$$

We will derive a hierarchical a posteriori error estimator for a given finite element discretization of this continuous problem. To this end let $(\mathcal{T}_0, \dots, \mathcal{T}_j)$ be a grid hierarchy obtained by successive local refinement of a conforming initial grid $\mathcal{T}_0$ and $\mathcal{S}(\mathcal{T})$ the space of conforming first-order finite element functions on the leaf grid $\mathcal{T} = \mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_j)$. Furthermore, we assume that elements are always refined such that new nodes are only introduced at the midpoints of adjacent edges. Analogously to Chapter 3 we use the discrete approximation

$$\psi^{\mathcal{T}}(v) = (\Psi(v), 1)^{\mathcal{T}} = \sum_{p \in \mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T})} \Psi(v(p)) \int_\Omega \lambda_p(x)\, dx$$

of the nonlinearity ψ on $\mathcal{S}(\mathcal{T})$ obtained by lumping. This leads to the discretization:

Problem 6.2. *Find $(u^{\mathcal{T}}, w^{\mathcal{T}}) \in \mathcal{S}(\mathcal{T}) \times \mathcal{S}(\mathcal{T})$ such that*

$$\begin{aligned} \left\langle \mathcal{F}(u^{\mathcal{T}}), v - u^{\mathcal{T}} \right\rangle + \psi^{\mathcal{T}}(v) - \psi^{\mathcal{T}}(u^{\mathcal{T}}) + b(v - u^{\mathcal{T}}, w^{\mathcal{T}}) &\geq \left(f, v - u^{\mathcal{T}}\right) \quad && \forall v \in \mathcal{S}(\mathcal{T}),\\ b(u^{\mathcal{T}}, v) - c(w^{\mathcal{T}}, v) &= (g, v) && \forall v \in \mathcal{S}(\mathcal{T}). \end{aligned}$$

While the space of conforming piecewise quadratic finite elements is often chosen as extended space for linear problems, it leads to instabilities if it is used for obstacle problems [54, 78]. For this reason we select the space of piecewise linear finite elements on a globally refined grid as extended space. More precisely let $\mathcal{T}'$ be a triangulation obtained by introducing new nodes at all edge midpoints and refining each element $\tau \in \mathcal{T}$ into 2^d new simplices.

Ideally we could solve the saddle point problem in the extended space $\mathcal{S}(\mathcal{T}')$ obtaining approximations $(u^{\mathcal{T}'}, w^{\mathcal{T}'})$ and estimate the error by

$$\|u^{\mathcal{T}} - u\|_1 + \|w^{\mathcal{T}} - w\|_1 \leq \|u^{\mathcal{T}} - u^{\mathcal{T}'}\|_1 + \|w^{\mathcal{T}} - w^{\mathcal{T}'}\|_1 + \|u - u^{\mathcal{T}'}\|_1 + \|w - w^{\mathcal{T}'}\|_1.$$

If the approximation in $\mathcal{S}(\mathcal{T}')$ is sufficiently better than the one in $\mathcal{S}(\mathcal{T})$ the term $\|u^{\mathcal{T}} - u^{\mathcal{T}'}\|_1 + \|w^{\mathcal{T}} - w^{\mathcal{T}'}\|_1$ can be used as error estimate. This condition is known as saturation assumption. Unfortunately the computation of $u^{\mathcal{T}'}, w^{\mathcal{T}'} \in \mathcal{S}(\mathcal{T}')$ is very expensive since it involves the solution of a nonlinear saddle point problem with the same algebraic structure but about 2^d times as many unknowns. In order to derive a cheaper estimate we first split the extended space by introducing a proper incremental space $\mathcal{V}$.

Theorem 6.1. *The space $\mathcal{Q} = \mathcal{S}(\mathcal{T}')$ can be split as $\mathcal{Q} = \mathcal{S}(\mathcal{T}) \oplus \mathcal{V}$ with the incremental space $\mathcal{V} := \operatorname{span} \mathcal{B}^{\mathcal{V}}$ for the linearly independent set $\mathcal{B}^{\mathcal{V}}$ defined by*

$$\mathcal{B}^{\mathcal{V}} := \left\{ \lambda_p^{\mathcal{T}'} \in \mathcal{B}(\mathcal{T}') \,|\, p \in \mathcal{N}^{\mathcal{V}} \right\}, \qquad \mathcal{N}^{\mathcal{V}} := \Big(\mathcal{N}(\mathcal{T}') \setminus \mathcal{H}(\mathcal{T}')\Big) \setminus \Big(\mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T})\Big).$$

Proof. First we note that, due to the restriction of new nodes to edge midpoints, $\mathcal{T}'$ is also the leaf grid of a grid hierarchy $(\mathcal{T}_0, \mathcal{T}_1', \mathcal{T}_2', \dots, \mathcal{T}_{j+1}')$ with $\mathcal{T}_i \subset \mathcal{T}_i'$ for $0 < i \leq j$. Hence by Lemma 3.4, Lemma 3.3, and Theorem 3.1 we have

$$\operatorname{span} \mathcal{B}(\mathcal{T}) = \mathcal{S}(\mathcal{T}) \subset \mathcal{S}(\mathcal{T}') = \operatorname{span} \mathcal{B}(\mathcal{T}').$$

Since $\mathcal{B}^{\mathcal{V}} \subset \mathcal{B}(\mathcal{T}')$ we know that $\mathcal{B}^{\mathcal{V}}$ is linear independent. Furthermore, by $\mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T}) \subset \mathcal{N}(\mathcal{T}') \setminus \mathcal{H}(\mathcal{T}')$ no $\lambda \in \mathcal{B}^{\mathcal{V}}$ can be spanned by $\mathcal{B}(\mathcal{T})$. Hence we have shown that $\mathcal{S}(\mathcal{T}) \oplus \mathcal{V} \subset \mathcal{S}(\mathcal{T}')$.

To see that we even have equality let $v \in \mathcal{S}(\mathcal{T}')$ and define $v^{\mathcal{T}} = I^{\mathcal{T}} v \in \mathcal{S}(\mathcal{T})$ as the linear interpolation in $\mathcal{S}(\mathcal{T})$ and $v^{\mathcal{V}} = v - v^{\mathcal{T}} \in \mathcal{S}(\mathcal{T}')$. Then by definition we have $v^{\mathcal{V}}(p) = 0$ for all $p \in [\mathcal{N}(\mathcal{T}') \setminus \mathcal{H}(\mathcal{T}')] \setminus \mathcal{N}^{\mathcal{V}}$ and thus $v^{\mathcal{V}} \in \mathcal{V}$. □

Note that for each $p \in \mathcal{N}^{\mathcal{V}}$ we either have $p \notin \mathcal{N}(\mathcal{T})$ or $p \in \mathcal{H}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T}')$. In the first case p is a new node in $\mathcal{T}'$ and thus placed at an edge mid point of $\mathcal{T}$. In the latter case p is a hanging node in $\mathcal{T}$ but no longer in $\mathcal{T}'$. Thus there must be an element $\tau \in \mathcal{T}$ such that p is on an edge of τ but a vertex of some element in $\{\tau' \in \mathcal{T}' \,|\, \tau' \subset \tau\}$ and thus also the midpoint of an edge in $\mathcal{T}$. Hence $\mathcal{N}^{\mathcal{V}}$ is the set of all edge midpoints in $\mathcal{T}$ that are not hanging nodes in $\mathcal{T}'$.

For the extension by piecewise quadratic finite elements the incremental space is given by the span of the so-called quadratic edge bubble-functions associated to the edges in $\mathcal{T}$. In $\mathcal{V}$ these are just replaced by the piecewise linear edge bubble functions associated to the edges in $\mathcal{T}$.

In order to simplify the problem that has to be solved for the error estimator we recall the procedure for a symmetric linear elliptic problem

$$u \in H^1(\Omega): \qquad a(u, v) = l(v) \qquad v \in H^1(\Omega). \tag{6.1}$$

Assume that $u^{\mathcal{T}}$ is the approximation in $\mathcal{S}(\mathcal{T})$ given by

$$u^{\mathcal{T}} \in \mathcal{S}(\mathcal{T}): \qquad a(u^{\mathcal{T}}, v) = l(v) \qquad v \in \mathcal{S}(\mathcal{T}).$$

Then we compute the solution $u^{\mathcal{Q}} = u^{\mathcal{T}} + (u^{\mathcal{Q}} - u^{\mathcal{T}})$ in the extended space by solving the defect problem

$$u^{\mathcal{Q}} \in \mathcal{Q}: \qquad a(u^{\mathcal{Q}} - u^{\mathcal{T}}, v) = l(v) - a(u^{\mathcal{T}}, v) \qquad v \in \mathcal{Q}.$$

The problem is localized by replacing the bilinear form $a(\cdot,\cdot)$ by a hierarchical preconditioner $a^{\mathcal{Q}} : \mathcal{Q} \times \mathcal{Q} \to \mathbb{R}$ given by

$$a^{\mathcal{Q}}(v,w) := a(I^{\mathcal{T}}v, I^{\mathcal{T}}w) + \sum_{p\in\mathcal{N}^{\mathcal{V}}} a(\lambda_p^{\mathcal{T}'}, \lambda_p^{\mathcal{T}'})(v - I^{\mathcal{T}}v)(p)(w - I^{\mathcal{T}}w)(p),$$

using the splitting $v = I^{\mathcal{T}}v + (v - I^{\mathcal{T}}v) \in \mathcal{Q}$ for $v, w \in \mathcal{Q}$ and the interpolation $I^{\mathcal{T}}$ introduced in Definition 3.9. Then we obtain an approximation $\tilde{u}^{\mathcal{Q}}$ of $u^{\mathcal{Q}}$ by solving the preconditioned defect problem

$$\tilde{u}^{\mathcal{Q}} \in \mathcal{Q} : \qquad a^{\mathcal{Q}}(\tilde{u}^{\mathcal{Q}} - u^{\mathcal{T}}, v) = l(v) - a(u^{\mathcal{T}}, v) \qquad v \in \mathcal{Q}.$$

Since $u^{\mathcal{T}}$ solves the variational equation in $\mathcal{S}$ we find that

$$a^{\mathcal{Q}}(\tilde{u}^{\mathcal{Q}} - u^{\mathcal{T}}, v) = a(I^{\mathcal{T}}(\tilde{u}^{\mathcal{Q}} - u^{\mathcal{T}}), v) = l(v) - a(u^{\mathcal{T}}, v) = 0 \qquad \forall v \in \mathcal{S}(\mathcal{T}).$$

Hence we have $I^{\mathcal{T}}(\tilde{u}^{\mathcal{Q}} - u^{\mathcal{T}}) = 0$ and $I^{\mathcal{T}}\tilde{u}^{\mathcal{Q}} = u^{\mathcal{T}}$. Conversely we have for any $w \in \mathcal{V}$

$$a^{\mathcal{Q}}(w, v) = 0 \qquad \forall v \in \mathcal{S}(\mathcal{T}).$$

Using this orthogonality we find that the increment $e^{\mathcal{V}} := \tilde{u}^{\mathcal{Q}} - I^{\mathcal{T}}\tilde{u}^{\mathcal{Q}} = \tilde{u}^{\mathcal{Q}} - u^{\mathcal{T}} \in \mathcal{V}$ is the solution of

$$e^{\mathcal{V}} \in \mathcal{V} : \qquad a^{\mathcal{V}}(e^{\mathcal{V}}, v) = l(v) - a(u^{\mathcal{T}}, v) \qquad \forall v \in \mathcal{V},$$

where $a^{\mathcal{V}} = a^{\mathcal{Q}}|_{\mathcal{V}\times\mathcal{V}}$ is the restriction of $a^{\mathcal{Q}}$ to the incremental space given by

$$a^{\mathcal{V}}(v,w) = \sum_{p\in\mathcal{N}^{\mathcal{V}}} a(\lambda_p^{\mathcal{T}'}, \lambda_p^{\mathcal{T}'})v(p)w(p).$$

Provided that the saturation assumption holds and that the hierarchical splitting is stable we have an approximation $e^{\mathcal{V}} \approx u^{\mathcal{Q}} - u^{\mathcal{T}} \approx u - u^{\mathcal{T}}$. Furthermore, the nodal values $e^{\mathcal{V}}(p)$ of $e^{\mathcal{V}}$ for $p \in \mathcal{N}^{\mathcal{V}}$ can be computed as solutions of local problems

$$e^{\mathcal{V}}(p) = \arg\min_{\rho\in\mathbb{R}} J(u^{\mathcal{T}} + \rho\lambda_p^{\mathcal{T}'}) = a(\lambda_p^{\mathcal{T}'}, \lambda_p^{\mathcal{T}'})^{-1}\Big(l(\lambda_p^{\mathcal{T}'}) - a(u^{\mathcal{T}}, \lambda_p^{\mathcal{T}'})\Big),$$

where $J(v) = \frac{1}{2}a(v,v) - l(v)$ is the energy associated with (6.1). Now the term

$$\|u^{\mathcal{V}}\|_{a^{\mathcal{Q}}} = \Big(\sum_{p\in\mathcal{N}^{\mathcal{V}}} \eta_p^2\Big)^{\frac{1}{2}}$$

with the local edge contributions

$$\eta_p = a(\lambda_p^{\mathcal{T}'}, \lambda_p^{\mathcal{T}'})^{\frac{1}{2}} e^{\mathcal{V}}(p) = \|e^{\mathcal{V}}(p)\lambda_p^{\mathcal{T}'}\|_a$$

for the norms $\|\cdot\|_{a^{\mathcal{Q}}}$ and $\|\cdot\|_a$ induced by $a^{\mathcal{Q}}$ and a can be used as an estimate for the global error. Consequently the local contributions η_p can be used as local error indicators for an adaptive strategy.

For nonlinear variational inequalities we can no longer use the hierarchical preconditioning and the orthogonality. However, we can generalize the approach to the nonlinear saddle point problem using local incremental problems directly. To this end we first define the restrictions $b^{\mathcal{V}}, c^{\mathcal{V}} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ of hierarchical preconditioners $b^{\mathcal{Q}}, c^{\mathcal{Q}}$ for the bilinear forms b and c analogously by

$$\begin{aligned} b^{\mathcal{V}}(v,w) &:= \sum_{p\in\mathcal{N}^{\mathcal{V}}} b(\lambda_p^{\mathcal{T}'}, \lambda_p^{\mathcal{T}'})v(p)w(p), \\ c^{\mathcal{V}}(v,w) &:= \sum_{p\in\mathcal{N}^{\mathcal{V}}} c(\lambda_p^{\mathcal{T}'}, \lambda_p^{\mathcal{T}'})v(p)w(p). \end{aligned}$$

For the nonlinear operator $\mathcal{F}$ we define the diagonalized shifted restriction $\mathcal{F}^{\mathcal{V}} : \mathcal{V} \to \mathcal{V}'$ to the incremental space by

$$\left\langle \mathcal{F}^{\mathcal{V}}(v), w \right\rangle := \sum_{p\in\mathcal{N}^{\mathcal{V}}} \left\langle \mathcal{F}(u^{\mathcal{T}} + v(p)\lambda_p^{\mathcal{T}'}), \lambda_p^{\mathcal{T}'} \right\rangle w(p).$$

Note that we cannot split the part depending on $u^{\mathcal{T}}$ and the increment for the nonlinear operator. Thus we shifted the nonlinearity by $u^{\mathcal{T}}$ before diagonalizing it on $\mathcal{V}$. For a linear operator $\langle \mathcal{F}(v), w\rangle = a(v,w)$ this definition reduces to

$$\left\langle \mathcal{F}^{\mathcal{V}}(v), w \right\rangle := a^{\mathcal{V}}(v,w) + a(u^{\mathcal{T}}, w).$$

Similarly we use a shifted version of the nonsmooth nonlinearity given by

$$\begin{aligned} \psi^{\mathcal{V}}(v) &:= \psi^{\mathcal{T}'}(u^{\mathcal{T}} + v) = \big(\Psi(u^{\mathcal{T}} + v), 1\big)^{\mathcal{T}'} \\ &= \sum_{p\in\mathcal{N}^{\mathcal{V}}} \Psi(u^{\mathcal{T}}(p) + v(p)) \int_\Omega \lambda_p^{\mathcal{T}'}(x)\,dx + \mathrm{const}(u^{\mathcal{T}}), \end{aligned}$$

where the last term is a constant depending on $u^{\mathcal{T}}$. Using these ingredients the incremental problem for the hierarchical estimator reads:

Problem 6.3. *Find* $(e_u^{\mathcal{V}}, e_w^{\mathcal{V}}) \in \mathcal{V} \times \mathcal{V}$ *such that*

$$\begin{aligned} \left\langle \mathcal{F}^{\mathcal{V}}(e_u^{\mathcal{V}}), v - e_u^{\mathcal{V}} \right\rangle + \psi^{\mathcal{V}}(v) - \psi^{\mathcal{V}}(e_u^{\mathcal{V}}) + b^{\mathcal{V}}(v - e_u^{\mathcal{V}}, e_w^{\mathcal{V}}) & \\ \geq \big(f, v - u^{\mathcal{T}}\big) - b(v - u^{\mathcal{T}}, w^{\mathcal{T}}) &\qquad \forall v \in \mathcal{V}, \\ b^{\mathcal{V}}(e_u^{\mathcal{V}}, v) - c^{\mathcal{V}}(e_w^{\mathcal{V}}, v) = (g, v) - b(u^{\mathcal{T}}, v) + c(w^{\mathcal{T}}, v) &\qquad \forall v \in \mathcal{V}. \end{aligned}$$

Testing the variational inequality with $v = e^{\mathcal{V}} + (\alpha - e_u^{\mathcal{V}}(p))\lambda_p^{\mathcal{T}'}$ and the variational equation with $\alpha\lambda_p^{\mathcal{T}'}$ for a fixed $p \in \mathcal{N}^{\mathcal{V}}$ we find that $(e_u^{\mathcal{V}}(p), e_w^{\mathcal{V}}(p)) \in \mathbb{R}^2$ can be computed as the solution of the local saddle point problem

$$\begin{aligned} F_p(e_u^{\mathcal{V}}(p))(\alpha - e_u^{\mathcal{V}}(p)) + \psi_p(\alpha) - \psi_p(e_u^{\mathcal{V}}(p)) + b_p(\alpha - e_u^{\mathcal{V}}(p))e_w^{\mathcal{V}} \geq f_p(\alpha - e_u^{\mathcal{V}}(p)) \quad \forall \alpha \in \mathbb{R} \\ b_p e_u^{\mathcal{V}} - c_p e_w^{\mathcal{V}} = g_p \end{aligned}$$

with the local representations

$$\begin{aligned} F_p(\alpha) &= \left\langle \mathcal{F}^{\mathcal{V}}(\alpha\lambda_p^{\mathcal{T}'}), \lambda_p^{\mathcal{T}'} \right\rangle, & \psi_p(\alpha) &= \Psi(u^{\mathcal{T}}(p) + \alpha\lambda_p^{\mathcal{T}'}) \int_\Omega \lambda_p^{\mathcal{T}'}(x)\,dx, \\ b_p &= b(\lambda_p^{\mathcal{T}}, \lambda_p^{\mathcal{T}}), & f_p &= \big(f, \lambda_p^{\mathcal{T}}\big) - b(\lambda_p^{\mathcal{T}}, w^{\mathcal{T}}), \\ c_p &= c(\lambda_p^{\mathcal{T}}, \lambda_p^{\mathcal{T}}), & g_p &= \big(g, \lambda_p^{\mathcal{T}}\big) - b(u^{\mathcal{T}}, \lambda_p^{\mathcal{T}}) + c(w^{\mathcal{T}}, \lambda_p^{\mathcal{T}}). \end{aligned}$$

Using the subdifferential $\partial\psi_p$ this can also be written as the two-dimensional problem

$$\begin{pmatrix} F_p + \partial\psi_p & b_p \\ b_p & -c_p \end{pmatrix} \begin{pmatrix} e_u^{\mathcal{V}}(p) \\ e_w^{\mathcal{V}}(p) \end{pmatrix} \ni \begin{pmatrix} f_p \\ g_p \end{pmatrix}.$$

In order to have a unique solution we assume that $c_p > 0$ or, equivalently, that no $\lambda_p^{\mathcal{T}'}$ is contained in the kernel of c. Then we can compute the solution by finding a $e_u^{\mathcal{V}}(p)$ such that

$$0 \in (F_p + \partial_p)(e_u^{\mathcal{V}}(p)) + \left(c_p^{-1} b_p^2\right) e_u^{\mathcal{V}}(p) - (f + c_p^{-1} g),$$

for example using the bisection method. There is a unique solution since $F_p + \partial\psi_p$ is maximal monotone and $c_p^{-1}b_p^2$ is positive. The other component can then be computed as

$$e_w^{\mathcal{V}}(p) = c_p^{-1}(b_p e_u^{\mathcal{V}}(p) - g_p).$$

Similar to the minimization formulation for the linear case presented above this local solution can be interpreted as the saddle point of the Lagrangian functional $L_p : \mathbb{R}^2 \to \mathbb{R} \cup \{\infty\}$ given by

$$L_p(e_u^{\mathcal{V}}(p), e_w^{\mathcal{V}}(p)) = L^{\mathcal{T}'}(u^{\mathcal{T}} + e_u^{\mathcal{V}}(p)\lambda_p^{\mathcal{T}'}, w^{\mathcal{T}} + e_w^{\mathcal{V}}(p)\lambda_p^{\mathcal{T}'})$$

where $L^{\mathcal{T}'} : \mathcal{Q} \times \mathcal{Q} \to \mathbb{R} \cup \{\infty\}$,

$$L^{\mathcal{T}'}(u, w) = J_0(u) + \psi^{\mathcal{T}'}(u) - (f, u) + b(u, w) - (g, w) - \frac{1}{2}c(w, w),$$

is the functional associated with Problem 6.1 except that the nonlinearity is approximated with respect to $\mathcal{Q}$.

Now an approximation of the global error $(\|u - u^{\mathcal{T}}\|_{\text{primal}}^2 + \|w - w^{\mathcal{T}}\|_{\text{dual}}^2)^{1/2}$ with properly scaled norms $\|\cdot\|_{\text{primal}}$ and $\|\cdot\|_{\text{dual}}$ for u and w on $H^1(\Omega)$ can be computed by

$$e(u^{\mathcal{T}}) = \Big(\sum_{p \in \mathcal{N}^{\mathcal{V}}} \eta_p^2\Big)^{\frac{1}{2}}, \qquad \eta_p^2 = \|e_u^{\mathcal{V}}(p)\lambda_p^{\mathcal{T}'}\|_{\text{primal}}^2 + \|e_w^{\mathcal{V}}(p)\lambda_p^{\mathcal{T}'}\|_{\text{dual}}^2.$$

Again the edge contributions η_p for the edge midpoints $p \in \mathcal{N}^{\mathcal{V}}$ can be used as local error indicator for an adaptive strategy. We will use the following strategy proposed by Dörfler [46] for the Poisson equation:

The intention is to refine a set of edges that "produce" a fixed fraction $\alpha \in (0,1]$ of the estimated global error $e(u^{\mathcal{T}})^2$. Moreover the set should be as small as possible. To this end define the sets

$$\mathcal{N}(\mathrm{Tol}) = \{p \in \mathcal{N}^{\mathcal{V}} \,:\, \eta_p^2 \geq \mathrm{Tol}\}$$

of all local indicators that are larger than a given tolerance Tol and compute

$$\mathrm{Tol}(\alpha) = \min\Big\{\eta_p^2 \,:\, p \in \mathcal{N}^{\mathcal{V}}, \sum_{q \in \mathcal{N}(\eta_p^2)} \eta_q^2 < \alpha e(u^{\mathcal{T}})^2\Big\}.$$

Then the set $\mathcal{N}(\mathrm{Tol}(\alpha))$ is the smallest set such that

$$\sum_{p \in \mathcal{N}(\mathrm{Tol}(\alpha))} \eta_p^2 < \alpha e(u^{\mathcal{T}})^2.$$

It can easily be obtained by sorting the local contributions by their values and then, starting from the largest value, including points p into $\mathcal{N}(\mathrm{Tol}(\alpha))$ as long as the sum of the local contributions in the set is smaller then the fraction α. Finally we mark all elements in $\tau \in \mathcal{T}$ such that $p \in \mathcal{N}(\mathrm{Tol}(\alpha))$ is the midpoint of an adjacent edge of τ.

This strategy was analyzed by Dörfler [46] for the approximate solution of the Poisson equation with adaptive linear finite elements on conforming grids. There it was shown that the produced sequence of finite element solutions converges to the weak solution of the continuous problem if the incremental spaces spanned by the quadratic edge bubble functions are used.

6.2 Implementation Aspects

In this section we discuss some implementation aspects that are crucial for the adaptive algorithm proposed in the previous section, especially if the solution of the stationary problem is a subproblem for one time step of a Rothe method.

6.2.1 Spatial Adaptivity for Rothe's Method

Consider the evolution equation

$$\left\langle \frac{du}{dt}, v-u \right\rangle + \langle \mathcal{F}(u), v-u\rangle + (g(u), v-u) \geq 0 \qquad \forall v \in H, \text{ a.e. in } (0,T],$$

with $u(0) = 0$ for a function $u : [0,T] \to H$ in some function space H on the domain Ω. Here $\mathcal{F} : H \to H'$ is a possibly set-valued nonlinear differential operator and $g : \mathbb{R} \to \mathbb{R}$ a scalar function. For example, linear parabolic equations as well as the Allen–Cahn and the Cahn–Hilliard equations presented in Chapter 2 can be expressed in this way.

We assume that Rothe's method is applied to this equation using the semi-implicit time discretization

$$\left(\frac{u_k - u_{k-1}}{\Delta t_k}, v-u_k\right) + \langle \mathcal{F}(u_k), v-u_k\rangle + (g(u_{k-1}), v-u_k) \geq 0 \qquad \forall v \in H, k > 0$$

where u_k denotes an approximation of $u(t_k)$. Again we can represent the semi-implicit and the fully implicit time discretizations introduced in Section 3.4 for the Cahn–Hilliard equation in this way.

Although we discretize the evolution problem in function space, we solve each stationary problem approximately in a discrete finite element space. The main advantage of Rothe's method is that different adaptive grids can be used in each time step. Let $\mathcal{T}_{k-1}$ be a locally refined grid and assume that the approximation from the previous time step is some $u_{k-1}^{\mathcal{T}_{k-1}} \in \mathcal{S}(\mathcal{T}_{k-1})$ in the finite element space $\mathcal{S}(\mathcal{T}_{k-1})$. Then the approximation for the current time step k is computed in $\mathcal{S}(\mathcal{T}_k)$ for an adaptive grid $\mathcal{T}_k$ by

$$\begin{aligned}\left(u_k^{\mathcal{T}_k}, v - u_k^{\mathcal{T}_k}\right) + \Delta t_k \left\langle \mathcal{F}^{\mathcal{T}_k}(u_k^{\mathcal{T}_k}), v - u_k^{\mathcal{T}_k} \right\rangle& \\ \geq \Delta t_k \left(\Delta t_k g(u_{k-1}^{\mathcal{T}_{k-1}}) + u_{k-1}^{\mathcal{T}_{k-1}}, v - u_k^{\mathcal{T}_k}\right) & \qquad \forall v \in \mathcal{S}(\mathcal{T}_k).\end{aligned}$$

In general, we will have $\mathcal{T}_{k-1} \neq \mathcal{T}_k$. Hence, expressions of the form

$$\left(w^{\mathcal{T}_{k-1}}, v^{\mathcal{T}_k}\right)$$

for grid functions $w^{\mathcal{T}_{k-1}} \in \mathcal{S}(\mathcal{T}_{k-1})$ and $v^{\mathcal{T}_k} \in \mathcal{S}(\mathcal{T}_k)$ have to be computed. Several strategies are used to deal with this problem:

1. Two separate grids objects can be used. In this case two problems appear: With typical data structures for finite element functions it is quite difficult to relate a function from one grid to elements from the other. Moreover, this approach has the drawback that storing two adaptive grid objects uses a lot of memory.

2. Only use a full grid object for the current time step and store nodal coordinates and values for the function $w^{\mathcal{T}_{k-1}}$ from the previous time step. This approach will introduce errors since the available information does not allow to reconstruct the exact function.

3. Construct the current adaptive grid by modifying the grid $\mathcal{T}_{k-1}$ from the previous time step. Starting from the grid $\mathcal{T}_{k-1}$ the grid for the time step k can be refined or coarsened where necessary during the adaptive cycle. Every time the grid is refined in a region, the solution from the previous time step can be interpolated to the new grid. In regions where the grid is coarsened the solution can be projected to the coarse grid.

   Although this approach is used often, it has several drawbacks: The projection needed for coarsening will in general introduce errors. Moreover, it is not possible to re-refine the grid where it was already coarsened since the old function (before projection) values cannot be reconstructed. Also, the adaptive loop needs a coarsening strategy in addition to a refinement strategy. Finally, this approach will in general be expensive because the problem size will in most cases not vary much from one time step to the next. This implies that in each time step several (at least two) problems of approximately the same size must be solved.

All these strategies are unsatisfactory for the reasons stated above. As a remedy the Dune-Subgrid module was developed [60] on the basis of the dune library. The next subsections will describe the Dune-Subgrid module and show how it allows to avoid these problems.

6.2.2 The Dune-Subgrid Module

Dune is an object oriented C++ library [11, 12] that provides interfaces for common functionality needed for grid based methods for partial differential equations. Most importantly it provides a unified interface to access and manipulate adaptive hierarchical grids. This interface is implemented by different grid managers. While the Dune library itself does only come with grid managers for structured grids, it includes bindings for the unstructured adaptive grid managers UG [10], ALUGrid [27, 99], and Alberta [97].

A consequence of this design is the possibility to implement so-called meta-grids that modify one or more other underlying grids. The fact that all grid implementations can be accessed through the same interface allows to write generic libraries and applications where the grid manager can be exchanged without changing the actual application code. Since these interfaces are realized using C++ template techniques they only lead to a very small runtime overhead [11].

The basic feature of the Dune-Subgrid module is a meta grid manager provided by the `SubGrid` class that allows to treat a subset of a grid hierarchy as a grid hierarchy in its own right. Let $(\mathcal{T}_0, \dots, \mathcal{T}_j)$ be a grid hierarchy provided by a Dune grid manager. We will call this hierarchy the "host grid". Then the tuple $(\mathcal{T}'_0, \dots, \mathcal{T}'_{j'})$ is called a "subgrid" if $\mathcal{T}'_k \subset \mathcal{T}_k$ for $k = 0, ..., j'$ and if the tuple is a grid hierarchy itself.

The `SubGrid` class allows to deal with such subgrids in an efficient way. An object of the `SubGrid` class is created by marking a subset of the host grid's elements for inclusion. The marks and consecutive indices for these elements and adjacent faces, edges, and nodes are stored internally. After creation the `SubGrid` allows to access the subgrid $(\mathcal{T}'_0, \dots, \mathcal{T}'_{j'})$ induced by the marks.

Note that the subgrid does not store the elements, their geometry, and their father and neighbor relation itself. If the user wants to iterate over a level $\mathcal{T}'_k$ of the subgrid it internally iterates over the whole host grid level $\mathcal{T}_k$ skipping the elements that are not contained in $\mathcal{T}'_k$. Similarly iterating over the subgrid leaf $\mathcal{T}' = \mathcal{L}(\mathcal{T}'_0, \dots, \mathcal{T}'_{j'})$ iterates over all host grid level grids $\mathcal{T}_k$ with $k \leq j'$ skipping the elements that are not contained in $\mathcal{T}'$. Thus storing a subgrid through the `SubGrid` class needs significantly less memory than storing a copy of it as a classical stand alone grid manager, at least if the subgrids size not too small compared to the host grids size.

Since the `SubGrid` class implements the standard Dune grid interface all algorithms that run with classical Dune grid managers can also be used on a subgrid. This does also include adaptive algorithms. If an element $\tau \in \mathcal{T}'$ of the subgrid is marked for refinement one of the following two things happens:

- If τ is not contained in the host grids leaf $\mathcal{T} = \mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_j)$, then the direct children of τ are included in the subgrid.

- If $\tau \in \mathcal{T}$ then the element is marked for refinement in the host grid. After refinement of the host grid the direct children of the element are included in the subgrid.

This strategy allows to transparently refine a subgrid such that it grows in the existing host grid where possible and refines the host grid only where needed.

One important consequence is that $\mathcal{T}'$ is in general nonconforming even if $\mathcal{T}$ is conforming, since elements introduced as "closure" during a red–green refinement in the host grid are not necessarily included in the subgrid. In order to guarantee a certain regularity, the `SubGrid` class allows to restrict the maximal level difference of elements in $\mathcal{T}'$ that share a vertex. If such a restriction is imposed additional elements are refined if necessary.

Besides the standard grid interface the `SubGrid` class also provides methods that relate the subgrid to the host grid. For example one can check whether a host grid element is contained in the subgrid. If this is the case one can construct the object representing the element as part of the subgrid from an object representing it with respect to the host grid and the other way around. For a more detailed survey of the `SubGrid` interface and the internal implementation we refer to [60].

6.2.3 Spatial Adaptivity for Rothe's Method using Dune-Subgrid

Subgrids allow to efficiently implement spatial adaptivity for time-dependent problems avoiding the problems mentioned in Section 6.2.1. We now describe how to achieve this and discuss its advantages and possible drawbacks afterwards.

Assume that u_{k-1} is the solution of the spatial problem from the previous time step as finite element function with respect to a leaf grid $\mathcal{T} = \mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_j)$ of a grid hierarchy $(\mathcal{T}_0, \dots, \mathcal{T}_j)$. For the current time step we want to solve a problem using an adaptive grid sequence that involves the integration of a product of u_{k-1} with a function from the current grid. This requires the evaluation of u_{k-1} on suitable quadrature points in the elements of the current grid. If we consider $(\mathcal{T}_0, \dots, \mathcal{T}_j)$ as a host grid the adaptive solution in the current time step can be done using the following algorithm.

1. Create an initial coarse subgrid $(\mathcal{T}_0^0, \dots, \mathcal{T}_{j_0}^0)$ of the host grid $\mathcal{T}$.

2. Solve the problem on the leaf of the current subgrid. If u_{k-1} needs to be evaluated on the subgrid, evaluate it on the host grids leaf and transfer the information to the first ancestor that is contained in the subgrid. This can be done using the father relation in the host grid and the subgrid–host grid relation provided by the `SubGrid` class.

3. Estimate the discretization error for the solution on the current subgrid. If the desired tolerance is matched, go to step 5.

4. Adaptively refine the subgrid. If the host grid was refined implicitly by the `SubGrid` class, transfer the function u_{k-1} to the new host grid leaf level using

standard techniques. This can be done exactly, since the host grid is only refined and never coarsened implicitly. Go to step 2.

5. Coarsen the host grid such that it is as coarse as possible while still containing the subgrid, and transfer the current solution from the subgrid to the new host grid. This can again be done exactly, since the host grid can never be coarser than the subgrid.

In all steps the host grid is fine enough to represent solutions from the previous time step and from the current one. Only after the coarsening in the last step, solutions from the previous time step can no longer be represented. This is, however, no problem, since they are no longer needed for the next time step. This approach has several advantages compared to the approaches presented in Section 6.2.1.

Compared to the first approach using two separate grid it is much more efficient, since the subgrid needs significantly less memory than a separate grid. Moreover, the evaluation of the function from the old grid on the current one can be implemented efficiently using the relation of the subgrid to the host grid.

In contrast to the second approach that stores nodal values and coordinates of the function from the old grid, the use of a subgrid allows for the exact evaluation of u_{k-1} on the current grid. This is especially important since many evolution problems conserve certain integrals, which can only be inherited by the discretization if the respective integration is done exactly.

The latter is also an advantage compared to the third approach that modifies the grid from the previous time step. Furthermore, the subgrid approach is in general much more efficient since it allows to build the grid for the current time step from a very coarse grid. Assuming that an optimal solver is available and that the problem size increases by a factor of at least 2 during refinement, the overall computational effort is about $2n$ where n is the problem size on the finest grid. If the grid from the previous time step needs to be modified k times to construct the current grid, this implies this amounts in an effort of $(k+1)n$.

Numerical experiments for the heat equation [60] have shown that the overall computational effort for the adaptive refinement loop is at most 2.5-times as large as the effort for the solution on the resulting fine grid. In most time steps the factor was even contained in $[1.5, 2]$ which is a real improvement compared to the factor $k+1$ where k is at least equal to one but in almost all cases larger.

The factor becomes even larger, if we also take into account that, for strongly nonlinear problems, solutions from the coarse grid do in general provide a better initial guess than solutions from the previous time step.

The drawbacks are the following. On the one hand the computation of subgrid indices takes additional time. On the other hand iterating over the subgrid is slower than iterating over the host grid for two reasons. Even if the subgrid level $\mathcal{T}'_k$ is much smaller than the host grid level $\mathcal{T}_k$, one must iterate over the whole set $\mathcal{T}_k$ internally. Moreover, the additional layer introduced by wrapping the host grid in the subgrid make each grid access slower. Computational experiments in [60] showed that this leads to an overhead of about 37% when assembling a stiffness matrix for the Laplace operator where the subgrid covers the whole host grid. In practice this is not a problem, because assembling the problem takes only a small fraction of time compared to the solution of the nonlinear saddle point problems we are interested in. Also, certain optimizations in the implementation may allow to reach a smaller number in the future.

7 Numerical Results

In this chapter we test the methods developed in the previous chapters numerically. Since we did neither prove error bounds for the finite element discretization nor reliability or efficiency of the error estimator we will not do extensive test for them. However, we will show that the local error indicators allow the construction of adaptive grids that capture local properties of the solution well.

The numerical tests of the algebraic solvers will concentrate on the Schur Nonsmooth Newton method applied to the Cahn–Hilliard equation. The TNNMG method will be used as solver for the nonlinear subproblems in each step of the Schur Nonsmooth Newton method. Hence its efficiency will directly influence the efficiency of the latter. Although both algebraic solvers are applicable to a variety of problems, testing them for other problems than the Cahn–Hilliard equation is not within the scope of this work.

For further numerical examples regarding the TNNMG method we refer to [58, 62, 94]. There is was applied to quadratic obstacle problems, contact problems in elasticity, and the Allen–Cahn equation, respectively. Neither of these problems includes a nonquadratic smooth energy so the anisotropic problems considered here are truly novel. For the Schur Nonsmooth Newton method we refer to [57, 59], where it was applied to the Cahn–Hilliard equation and a control problem for the Poisson equation with control constraints. In both cases the convex energy of the saddle point problem was a quadratic energy with obstacles. Thus the presented anisotropic problem with logarithmic potential is again more general.

All numerical examples presented in the following use the semi-implicit time discretization of the Cahn–Hilliard equation given by Problem 3.6 with uniform time step size $\Delta t_k = \Delta t$ and $\theta_C = 1$. For the rank-1 regularization in (3.6) and the time step size we select $\rho = \overline{\gamma}$ and $\Delta t = 2\overline{\gamma}$, respectively. The norms used for the phase field and the chemical potential are

$$\begin{aligned} \|u\|_{\text{primal}}^2 &= \overline{\gamma}\,(\nabla u, \nabla u) + \overline{\gamma}\,(u,1)\,(u,1)\,, \\ \|w\|_{\text{dual}}^2 &= \Delta t\,(\nabla w, \nabla w) + \Delta t\,(w,w)\,. \end{aligned}$$

We will present results for the isotropic Cahn–Hilliard equation with

$$\gamma(x)^2 = \overline{\gamma}\|x\|^2$$

and for the anisotropic Cahn–Hilliard equation with a scaled version of the Kobayashi anisotropy of Example 2.1. More precisely, we use

$$\gamma(x)^2 = \overline{\gamma}\Big(1 + \bar{a}\cos(k\beta(x))\Big)^2\|x\|^2,$$

with $k = 3$ and $\bar{a} = 0.124 < 1/(k^2 - 1)$, where $\beta(\xi) \in [0, 2\pi]$ denotes the angle between the positive x-axis and ξ.

The spatial discretization is done using linear finite elements leading to the discrete saddle point problems in Problem 3.12. The equations are discretized on the domain $\Omega = (-1, 1)^2$ with a symmetric coarse grid consisting of four triangles, each spanned by the origin and two vertices of $\partial\Omega$.

For the solution of the algebraic problems we use the Schur Nonsmooth Newton method according to Corollary 5.3 with the following parameters:

- The truncated index set $\mathcal{I}''(v)$ uses the truncation criterion
$$\varphi_i''(v_i) < (C_\varphi)_{i,i} = 10^6 \int_\Omega \lambda_{p_1}(x)\, dx.$$
- We do not apply any additional correction $\mathcal{C}$, i.e., we use $\mathcal{C} = 0$.
- As step size rule we use the inexact step sizes introduced in Proposition 5.4 with the accuracy $\epsilon = 0.8$.
- The step size rule is switched off dynamically according to the strategy in Theorem 5.4 if the norm of the direction contracts with $\sigma = 0.8$. We select $\alpha_{-1} > 0$ such that the step rule is not used for the first correction.
- The linear saddle point problems are solved using a linear multigrid method with a Vanka-type block Gauß–Seidel smoother that solves local 2×2 saddle point problems for all vertices successively. The multigrid method uses a V-cycle with 3 pre- and post-smoothing steps. For the generated sequence $d^{\nu,\mu} \to d^\nu$ the relative correction norm
$$\frac{\|d^{\nu,\mu} - d^{\nu,\mu-1}\|_{\text{dual}}}{\|d^{\nu.\mu}\|_{\text{dual}}}$$
is used as estimate of the relative error.
- The linear saddle point problems are solved inexactly such that the estimated relative error is bounded by
$$\max\{0.1^\nu, e_{\nu-1}^2\}$$
where $e_{\nu-1} = \|w^{\nu-1} - w^{\nu-2}\|_{\text{dual}}$ is an estimate for the error of the previous step of the Schur Nonsmooth Newton method.
- The convex nonlinear minimization problems for the evaluation of F^{-1} are solved using the TNNMG method.

The TNNMG method is used with the following parameters:

- As nonlinear smoother we use the inexact nonlinear Gauß–Seidel method according to Theorem 4.2 with $\omega_0 = 0.5$.

- Three pre-smoothing steps of the nonlinear smoother are applied, followed by a truncated linear coarse correction and another three nonlinear smoothing steps.
- The truncated index set $\mathcal{I}''(v)$ uses the truncation criterion

$$\varphi_i''(v_i) < (C_\varphi)_{i,i} = 10^6 \int_\Omega \lambda_{p_1}(x)\, dx.$$

- The truncated linear system is solved inexactly with the multigrid method described in Section 4.3.1 using one V-cycle with 3 pre- and post-smoothing steps.

The implementation used for the numerical tests is based on the Dune framework. The following Dune modules where used:

- The core modules Dune-Common, Dune-Grid, Dune-Localfunctions, and Dune-Istl, where used for the grid interface, shape functions, and matrix and vector classes.
- The Dune-Subgrid module was used for the handling of adaptive grids during Rothe's method.
- The discretization module Dune-Fufem was used for the handling of grid functions and finite element spaces, the assembling of matrices and vectors, and for the implementation of the hierarchical error estimator.
- The solver module Dune-Solvers was used as infrastructure for algebraic solvers, and the handling of multigrid transfer operators.
- The TNNMG method was implemented in the Dune-Tnnmg module.
- The Schur Nonsmooth Newton method and the nonlinearities where implemented in the Phase-Field module.

7.1 Cahn–Hilliard Equations on Uniform Grids

In order to allow for a better comparison of the solvers behavior for different problems we consider the Cahn–Hilliard equation for fixed $\overline{\gamma} = 4 \cdot 10^{-4}$ on a sequence of uniformly refined grids. The finest grid is obtained by eight refinements and contains 131 585 vertices.

7.1.1 Iteration History on a Fixed Grid

As a first test we investigate the convergence for the first time step of the Cahn–Hilliard equation with the initial value

$$u_0(x) = \begin{cases} 1 & \text{if } |x| \leq 0.5, \\ -1 & \text{else.} \end{cases}$$

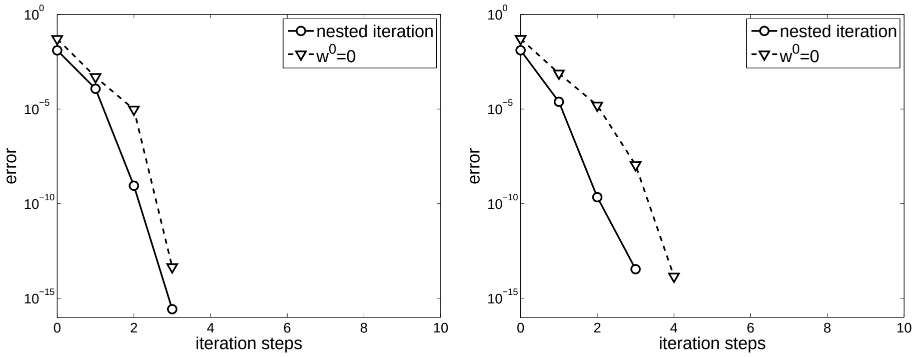


Figure 7.1: Circle example: Error over iteration step for the isotropic case (left) and the Kobayashi anisotropy (right).

This will be referred to as the “circle example” in the following.

Figure 7.1 depicts the algebraic error over iteration steps for the temperature $\theta = 0$, i.e., the obstacle potential, and both choices of γ. The error is approximated by $\|w^\nu - \tilde{w}\|_{\text{dual}}$ where $\tilde{w}$ is precomputed up to a significantly higher accuracy. The left and right plot show the iteration history for the isotropic case and the Kobayashi anisotropy. For both versions the dashed line corresponds to the initial guess $w^0 = 0$, while the solid line results from nested iteration, using the solution from a coarser level as initial guess. In any case superlinear convergence is observed and the iteration history seems not to depend strongly on the initial guess. Furthermore, the algorithm behaves the same for the isotropic and the anisotropy problem.

The same quantities are depicted in Figure 7.2 for a different initial value u_0 for the evolution. Here it is taken to be 1 inside of finitely many random circles and -1 outside. In contrast to the initial value in the circle example the transition between both phases is not a jump but smoothed to a sine-profile to model a typical situation during the phase transition evolution. This will be referred to as the “random discs example” in the following. Again the isotropic and the anisotropic problem both exhibit comparable superlinear convergence. However, nested iteration leads to faster convergence here.

7.1.2 Mesh Dependency

Now we investigate if and how the convergence speed of the algorithm changes if the grid is refined. To this end we consider the random discs example on different refinement levels and for the temperature $\theta = 0$. Since averaged convergence rates are inappropriate due to the superlinear convergence, we plot the number of Schur Nonsmooth Newton steps needed to achieve an algebraic error less than 10^{-13}.

The left and right plot in Figure 7.3 depict this number for the isotropic and the anisotropic case. Again we compare the initial guess $w^0 = 0$ and nested iteration. These examples show that the number of iteration steps does not increase for finer grids. Moreover, the solution on coarse grids does even take much more iteration steps.

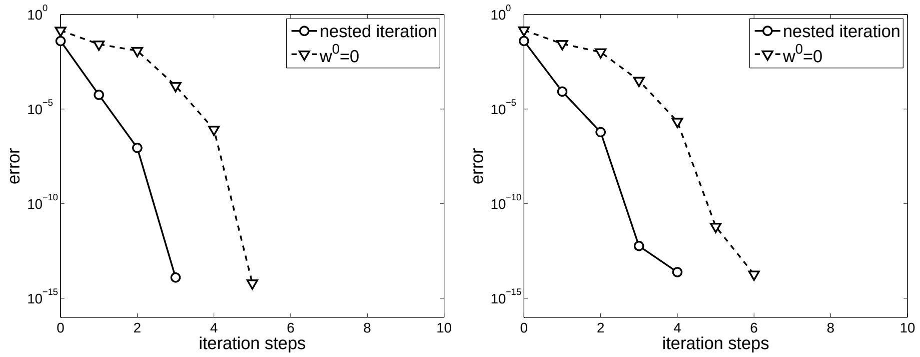


Figure 7.2: Random discs example: Error over iteration step for the isotropic case (left) and the Kobayashi anisotropy (right).

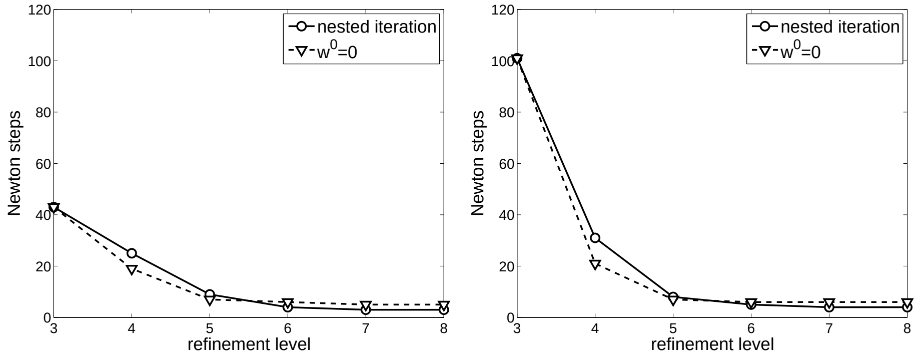


Figure 7.3: Random discs example: Newton steps over refinement level for the isotropic case (left) and the Kobayashi anisotropy (right).

This can be explained by the fact that very coarse grids cannot resolve the interface layer. Thus the inactive sets selected during the algorithm tend to be unstable. While the number of iteration steps is similar for fine grids, the anisotropic problem takes significantly more iteration steps on coarse grids compared to the isotropic one.

Until now we have only looked at the number of Schur Nonsmooth Newton steps. However, the computational work done in these iteration steps differs a lot for the following reasons. On the one hand the linear saddle point problems are solved with increasing accuracy leading to an increasing number of linear multigrid steps. On the other hand the operator F^{-1} has to be evaluated at least once in all steps, but several times if the step size rule is activated. Since the nonlinear multigrid steps are much more expensive than the linear ones we use the former as measure for the computational effort from now on.

For the examples of Figure 7.3 the numbers of overall TNNMG steps per refinement level are depicted in Figure 7.4. Most of the iterates on coarser grids require significant

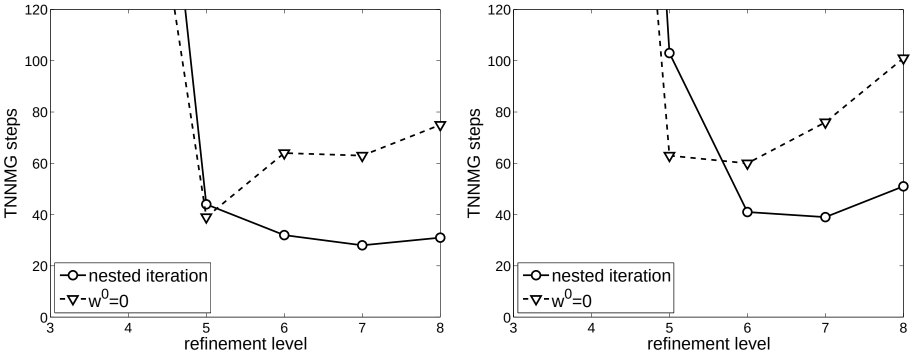


Figure 7.4: Random discs example: Overall TNNMG steps over refinement level for the isotropic case (left) and the Kobayashi anisotropy (right).

damping. Thus the number of TNNMG iterations does even take values larger than 1000 for the coarsest level. In order to provide a useful visualize for the more important finer grids we truncated the plot at 120. For these finer grids the number of TNNMG steps is bounded by 51 independently of the mesh size if nested iteration is applied. However, it takes about twice as much steps for the initial guess $w^0 = 0$.

In the light of the bounded number of outer iterations presented in Figure 7.3 this shows that nested iteration provides initial guesses that are particularly suited for the multigrid subproblem solvers, allowing for a mesh independent complexity of the overall algorithm. Due to this fact we will use nested iteration for all following examples.

7.1.3 Robustness with Respect to the Temperature

To examine if the convergence is also robust with respect to the temperature θ in the logarithmic potential, we now use the same example with θ varying from 10^{-9} to 10^2. The limiting case $\theta = 0$ was already discussed.

Figure 7.5 depicts the number of Schur Nonsmooth Newton steps needed on the eighth level to achieve an algebraic error less than 10^{-13} for different values of θ. The computational examples show that the convergence of the Schur Nonsmooth Newton method is hardly effected by the temperature.

Figure 7.6 depicts the number of overall TNNMG steps for the same example. Again the method is robust for $\theta \to 0$. Moreover, we observe faster convergence for high temperatures $\theta > 0$. This is not surprising since the nonlinearity dominates the differential operator in the Ginzburg–Landau energy in this case and the problems solved by the TNNMG method become “asymptotically diagonal” for $\theta \to \infty$.

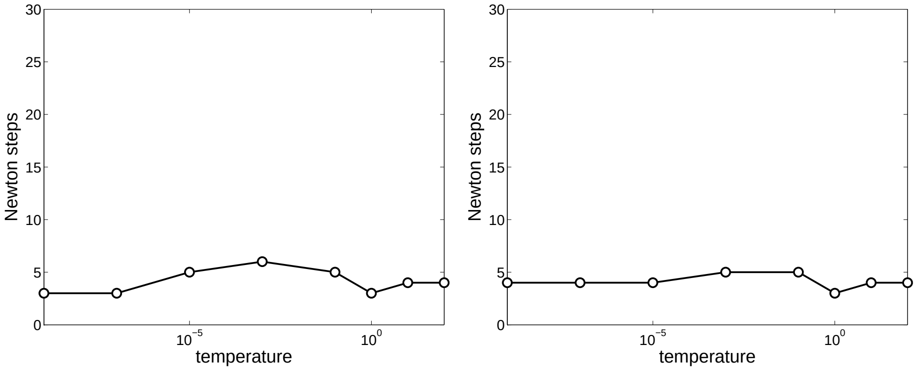


Figure 7.5: Random discs example: Newton steps over temperature for the isotropic case (left) and the Kobayashi anisotropy (right).

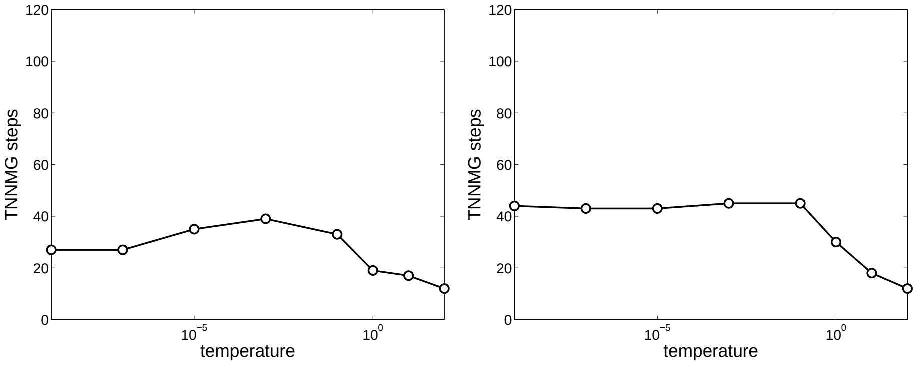


Figure 7.6: Random discs example: Overall TNNMG steps over temperature for the isotropic case (left) and the Kobayashi anisotropy (right).

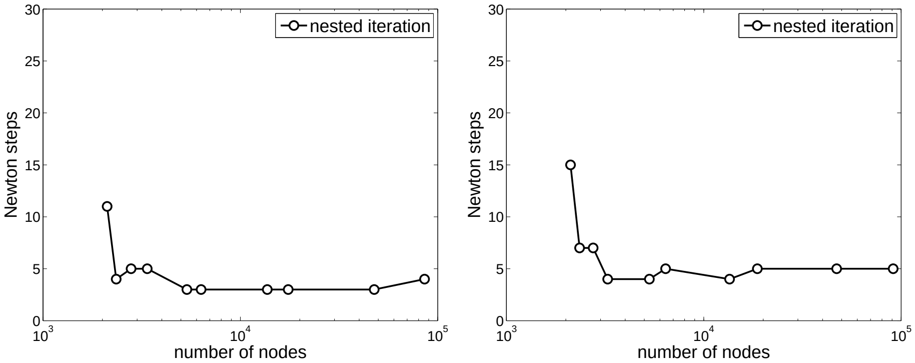


Figure 7.7: Circle example: Newton steps over number of unknowns for the isotropic case (left) and the Kobayashi anisotropy (right).

7.2 Cahn–Hilliard Equations on Adaptive Grids

Now we consider adaptive grids obtained using the local error indicators and the refinement strategy introduced in Section 6.1.

7.2.1 Mesh Dependency

We start with the circle example and use the smaller interface parameter $\overline{\gamma} = 10^{-4}$ now. Since the solution was most expensive for $\theta = 10^{-3}$ we consider this temperature for the tests presented here.

The adaptive grid is constructed starting from a grid obtained by five uniform refinements of the coarse grid. The hierarchical estimator uses the scaled norms introduced above and the refinement strategy is applied with the fraction $\alpha = 0.64$ as suggested by Dörfler [46] (more precisely $\theta^* = 0.2$ for $\theta^* = 1 - \sqrt{\alpha}$ was suggested there).

Figure 7.7 and Figure 7.8 depict the number of Newton iteration steps and the number of TNNMG steps over the number of nodes for the isotropic and the anisotropic interfacial energy. Similar to the uniform case it can be seen that the number of Newton steps does not increase with the number of nodes. While the same is true for the TNNMG steps in the isotropic case, the number of TNNMG steps slightly increases for the anisotropic case. However, it is still below 70 on the finest grid containing about 100 000 nodes. In any case the number is significantly higher for very coarse grids.

The same results are depicted in Figure 7.9 and Figure 7.10 for the random discs example. Here the number of iteration steps is bounded in all cases for an increasing number of nodes.

7.2.2 Hierarchical Error Estimator

In order to examine the error estimator and the quality of the resulting grids, we consider the same examples as in the previous section again.

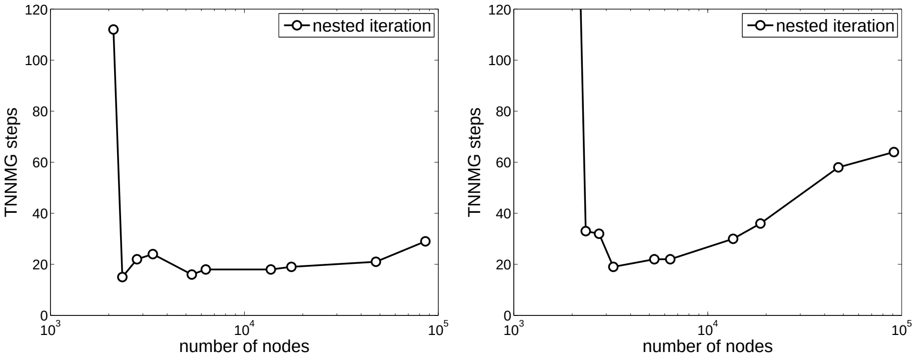


Figure 7.8: Circle example: Overall TNNMG steps over temperature for the isotropic case (left) and the Kobayashi anisotropy (right).

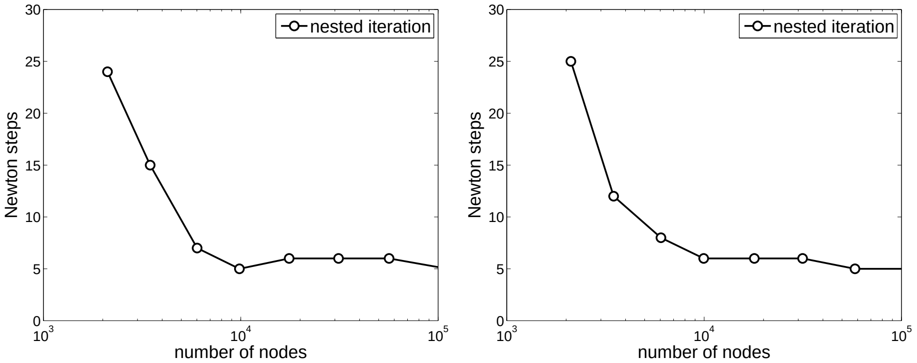


Figure 7.9: Random discs example: Newton steps over number of unknowns for the isotropic case (left) and the Kobayashi anisotropy (right).

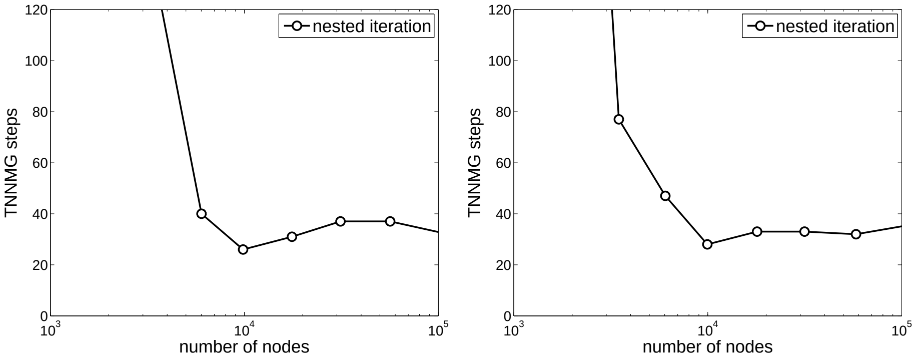


Figure 7.10: Random discs example: Overall TNNMG steps over temperature for the isotropic case (left) and the Kobayashi anisotropy (right).

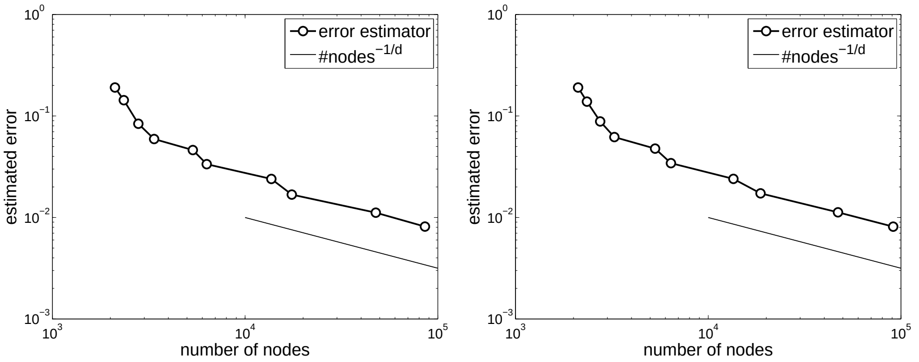


Figure 7.11: Circle example: Estimated error number of unknowns for the isotropic case (left) and the Kobayashi anisotropy (right).

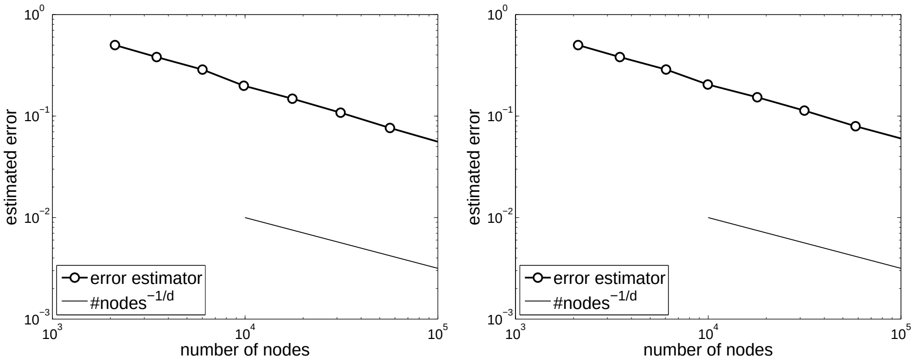


Figure 7.12: Random discs example: Estimated error number of unknowns for the isotropic case (left) and the Kobayashi anisotropy (right).

Figure 7.11 depicts the estimated error for the circle example and the isotropic and the anisotropic case, respectively. In order to simplify the interpretation a line with the slope of $\#\text{nodes}^{-1/d}$ that corresponds to the error order $O(h)$ for uniform grids is also plotted. After an initial phase, where the estimator decreases quickly, it takes almost exactly the slope of this line for the isotropic and the anisotropic case. For the random discs example in Figure 7.12 the slope can be seen even for coarse grids.

Finally we solve the anisotropic random discs example adaptively for $\overline{\gamma} = 10^{-3}$ and a tolerance of 0.1 for the adaptive refinement. Figure 7.13 shows the solution of the first time step and the adaptive grid. The closeup of the grid near the interface shows that the refinement strategy using the hierarchical estimator allows for grids that capture the interface very well. Figures 7.14 and 7.15 show further time steps of this problem. The threefold anisotropy is clearly visible from the shape of the initially round interface regions.

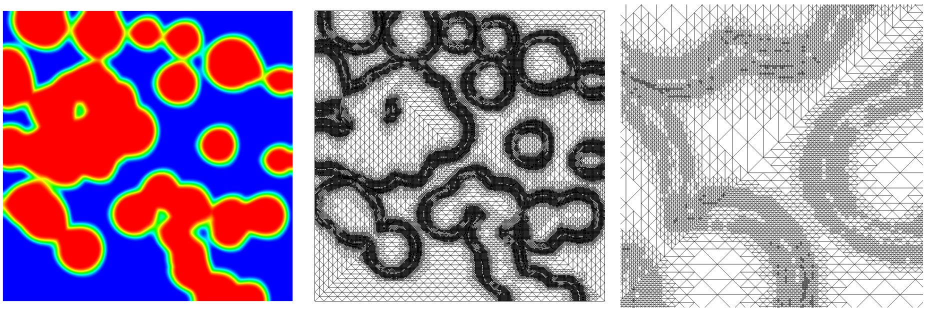


Figure 7.13: First time step: Solution (left), leaf grid (middle), and closeup of leaf grid (right).

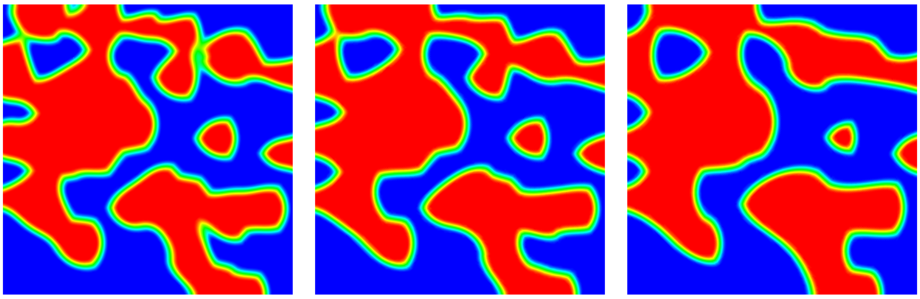


Figure 7.14: Time steps u_5 (left), u_{10} (middle), u_{40} (right).

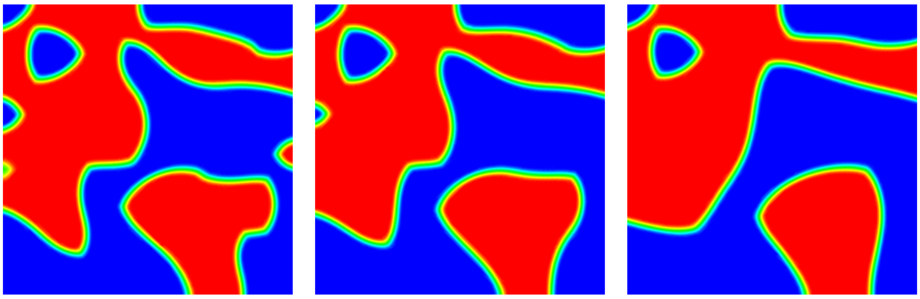


Figure 7.15: Time steps u_{80} (left), u_{120} (middle), u_{240} (right).

7.3 Conclusions

The numerical examples show that the Schur Nonsmooth Newton method converges independently on the mesh size in the sense that the number of iteration steps, needed to solve up to a fixed accuracy is bounded from above. Even more this number is in the range of 3 to 7 for fine grids. Only on very coarse grids, which can not resolve the interface layer, the number increases significantly. This behavior can be seen on uniform as well as on adaptively refined grids.

Furthermore, the experiments show that the number of iteration steps for a fixed accuracy is hardly influenced by the temperature. This result reflect the fact that the method was designed without any smoothness assumption on the decoupling nonlinearity. Instead of this convex properties that are shared by the obstacle potential and the logarithmic potential for all $\theta \geq 0$ where used.

If we use the number of TNNMG steps as measure for the computational effort, the situation is similar. However, the number of TNNMG steps is in most examples significantly smaller if nested iteration is used to compute the initial guess for the Schur Nonsmooth Newton method. In the light of the discussion in Section 4.2.5 this can be explained by the fact that the method is in general not mesh independent. Before fast multigrid convergence is attained the method needs an increasing number of iteration steps to determine the active set. In case of nested iteration the initial guess does typically also provide a good guess for the active set. Hence this pre-asymptotic phase is hardly visible in this case.

Finally we note that the hierarchical error estimator allows to construct grids that are perfectly matched to the interface. While similar grids could also be obtained using heuristic strategies that refine only near the interface, such strategies do in general require the selection of several parameters. Furthermore, they ignore regions where the solution is almost constant as they appear for computations with the logarithmic potential.

A Appendix

A.1 Properties of Convex Functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is called strongly convex with respect to a constant $\mu > 0$ and a semi-norm $\| \cdot \|_M^2 = \langle M\cdot, \cdot \rangle$ (for a symmetric positive semidefinite matrix M) if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \lambda(1-\lambda)\frac{\mu}{2}\|x-y\|_M^2 \qquad \forall \lambda \in [0,1] \quad \text{(A.1)}$$

holds true for all $x, y \in \mathbb{R}^n$.

Lemma A.1. *A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with respect to $\mu > 0$ and $\| \cdot \|_M$ if and only if*

$$f(x) - f(y) \geq \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x-y\|_M^2 \qquad \forall x, y \in \mathbb{R}^m \qquad \text{(A.2)}$$

or if ∇f is strongly monotone, i.e.

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x-y\|_M^2 \qquad \forall x, y \in \mathbb{R}^m.$$

Proof. Use literally the proof in [53] for the Euclidean norm. □

Theorem A.1. *Let V be a Banach space, $f_1, f_2 : V \to \mathbb{R} \cup \{\infty\}$ convex, $\lambda > 0$, and $x \in V$. Then*

1. $\partial(\lambda f_1)(x) = \lambda \partial f_1(x)$.
2. $\partial f_1(x) + \partial f_2(x) \subset \partial(f_1 + f_2)(x)$ *if f_1 and f_2 do not take the value ∞.*
3. $\partial f_1(x) + \partial f_2(x) = \partial(f_1 + f_2)(x)$ *if f_1 and f_2 are lower semicontinuous, not identical to ∞, and if there is some $x_0 \in \operatorname{dom} f_1 \cap \operatorname{dom} f_2$ where f_1 and f_2 are continuous.*

Proof. See [49, Chapter I, Section 5.3]. □

A.2 Properties of Function Spaces

Theorem A.2. *Let $\Omega \subset \mathbb{R}^d$ be a bounded, open set such that $\partial\Omega$ satisfies the uniform cone condition (cf. [110]). Then there is a constant $C > 0$ depending only on Ω such that*

$$\|v\|_1^2 \leq C_P \left(|v|_1^2 + (v,1)^2 \right) \leq C_P \left(|v|_1 + |\,(v,1)\,| \right)^2.$$

Proof. See [110]. □

Note that the uniform cone condition in Theorem A.2 is especially satisfied by a Lipschitz boundary.

The following result on superposition operators gives provides continuity of such operators under very few assumptions. For a comprehensive study on superposition operators we refer to [3].

Definition A.1. *Let $\Omega \subset \mathbb{R}^d$ be an open set and $f : \Omega \times \mathbb{R} \to \mathbb{R}$. The superposition operator F induced by f is given by*

$$(F(v))(x) = f(x, v(x)) \qquad \forall x \in \Omega$$

for functions $v : \Omega \to \mathbb{R}$. f is called a Carathéodory function if $f(x, \cdot)$ is continuous on $\mathbb{R}$ for almost all $x \in \Omega$ and if $f(\cdot, y)$ is measurable on Ω for all $y \in \mathbb{R}$.

Theorem A.3. *Let $\Omega \subset \mathbb{R}^d$ be an open set, $f : \Omega \times \mathbb{R} \to \mathbb{R}$ a Carathéodory function, and let the induced superposition operator F map L^p to L^q for some $1 \le p, q < \infty$. Then F is continuous.*

Proof. See [3]. □

Corollary A.1. *Let $\Omega \subset \mathbb{R}^d$ be a bounded, open set and $f : \mathbb{R} \to \mathbb{R}$ be continuous and bounded. Then the superposition operator F given by*

$$(F(v))(x) = f(v(x)) \qquad \forall x \in \Omega$$

is continuous from L^p to L^q for all $1 \le p, q < \infty$.

A.3 An Existence Result for Saddle Point Problems

Theorem A.4. *Let V, Z be Banach spaces, $\mathcal{A} \subset V$, $\mathcal{B} \subset Z$ convex, closed and nonempty sets. Let $L : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ satisfy*

$$\begin{aligned} &\forall \mu \in \mathcal{B}, \qquad && L(\cdot, \mu) : \mathcal{A} \to \mathbb{R} \text{ is convex and lower semicontinuous,} \\ &\forall v \in \mathcal{A}, \qquad && L(v, \cdot) : \mathcal{B} \to \mathbb{R} \text{ is concave and upper semicontinuous.} \end{aligned}$$

Furthermore, assume that

$$\begin{aligned} &\mathcal{A} \text{ is bounded} \qquad && \text{or} \qquad && \exists \mu_0 \in \mathcal{B} : \lim_{\substack{\|v\|_V \to \infty \\ v \in \mathcal{A}}} L(v, \mu_0) = \infty, \\ &\mathcal{B} \text{ is bounded} \qquad && \text{or} \qquad && \lim_{\substack{\|\mu\|_Z \to \infty \\ \mu \in \mathcal{B}}} \inf_{v \in \mathcal{A}} L(v, \mu) = -\infty. \end{aligned}$$

Then L possesses a saddle point $(u, \lambda) \in \mathcal{A} \times \mathcal{B}$, i.e.

$$L(u, \mu) \le L(u, \lambda) \le L(v, \lambda) \qquad \forall (v, \mu) \in \mathcal{A} \times \mathcal{B}.$$

Proof. See [49, Chapter VI, Proposition 2.4]. □

A.4 Convexification of a Class of Saddle Point Problems

Let H be a Hilbert space with norm $\|\cdot\|_H$. For functionals $f, g \in H'$, symmetric, positive semidefinite, linear operators $M, C : H \to H'$, a nonlinear operator $F : H \to H'$, a proper, convex, lower semicontinuous function $\varphi : H \to \mathbb{R} \cup \{\infty\}$ and $\tau > 0$ consider the saddle point problem:

Problem A.1. *Find $(u, w) \in H \times H$ such that*

$$\begin{aligned} \langle F(u), v-u\rangle + \varphi(v) - \varphi(u) - \langle Mw, v-u\rangle &\geq \langle f, v-u\rangle && \forall v \in H, \\ -\langle Mu, v\rangle - \tau\langle Cw, v\rangle &= \langle g, v\rangle && \forall v \in H. \end{aligned}$$

It will be help full to consider the modified problem:

Problem A.2. *Find $(u, \tilde{w}) \in H \times H$ such that*

$$\begin{aligned} \left\langle \tilde{F}(u), v-u\right\rangle + \varphi(v) - \varphi(u) - \left\langle \tilde{M}\tilde{w}, v-u\right\rangle &\geq \left\langle \tilde{f}, v-u\right\rangle && \forall v \in H, \\ -\left\langle \tilde{M}u, v\right\rangle - \tau\langle C\tilde{w}, v\rangle &= \langle g, v\rangle && \forall v \in H. \end{aligned}$$

Theorem A.5. *Let $\alpha > 0$. Then Problem A.1 is equivalent to Problem A.2 with*

$$\tilde{F}u = F(u) + 2\alpha M - \tau\alpha^2 C, \qquad \tilde{M} = M - \tau\alpha C, \qquad \tilde{f} = f - \alpha g,$$

in the sense that $(u, w) \in H \times H$ is a solution to Problem A.1. if and only if $(u, \tilde{w}) \in H \times H$ with $\tilde{w} = w + \alpha u$ is a solution to Problem A.2.

Proof. Adding $0 = \langle \alpha Mu - M\alpha u, v-u\rangle$ to the first and $0 = \langle \tau\alpha Cu - \tau C\alpha u, v\rangle$ to the second equation in Problem A.1 leads to

$$\begin{aligned} \langle F(u) + \alpha Mu, v-u\rangle + \varphi(v) - \varphi(u) - \langle M\tilde{w}, v-u\rangle &\geq \langle f, v-u\rangle && \forall v \in H, \\ -\left\langle \tilde{M}u, v\right\rangle - \tau\langle C\tilde{w}, v\rangle &= \langle g, v\rangle && \forall v \in H. \end{aligned}$$

Testing the second equation with $v - u$ and multiplying it by $-\alpha$ gives

$$\left\langle \left(\alpha M - \tau\alpha^2 C\right)u, v-u\right\rangle - \langle -\tau\alpha C\tilde{w}, v-u\rangle = -\alpha\langle g, v-u\rangle \qquad \forall v \in H.$$

Adding this to the first equation provides Problem A.2. Since we can revert all operations we have shown equivalence. □

Lemma A.2. *Let $F = \nabla J : H \to H'$ for a differentiable function $J : H \to \mathbb{R}$ and $|\cdot|_T^2 := \langle T\cdot, \cdot\rangle$ the seminorm induced by an operator T. Consider the Lagrange-functionals*

$$\begin{aligned} L(u, w) &= J(u) + \varphi(u) - \langle f, u\rangle + \langle -Mu - g, w\rangle - \frac{\tau}{2}|w|_C^2, \\ \tilde{L}(u, \tilde{w}) &= J(u) + \varphi(u) + \alpha|u|_M^2 - \alpha^2\frac{\tau}{2}|u|_C^2 - \langle f - \alpha g, u\rangle \\ &\qquad + \langle (\tau\alpha C - M)u - g, \tilde{w}\rangle - \frac{\tau}{2}|\tilde{w}|_C^2, \end{aligned}$$

associated with Problem A.1 and Problem A.2, respectively. Then $L(u, w) = \tilde{L}(u, w + \alpha u)$ and equivalently $L(u, \tilde{w} - \alpha u) = \tilde{L}(u, \tilde{w})$.

If F is not monotone itself Theorem A.5 allows to state an equivalent saddle point problem where $\tilde{F}$ might be monotone. More precisely, if F is the gradient of some functional it allows to add some convex quadratic term with respect to M at the price of adding a concave quadratic term with respect to C:

Theorem A.6. *Let $C_0 : H \to H'$ be a linear, continuous, symmetric, positive semidefinite operator such that $C + C_0$ is coercive, i.e. $|\cdot|_{C+C_0}$ is equivalent to the norm $\|\cdot\|_H$ of H. Assume that $F = \nabla J - 2\alpha M$ with $\alpha > 0$ and a convex, differentiable functional $J : H \to \mathbb{R}$ that is also strongly convex or equivalently ∇J is strongly monotone, i.e. there is a constant $\overline{\gamma} > 0$ with*

$$\langle \nabla J(u) - \nabla J(v), u - v \rangle \geq \overline{\gamma} |u - v|^2_{C+C_0}.$$

Then the operator $\tilde{F}$ from Theorem A.5 is strongly monotone if $\tau < \frac{\overline{\gamma}}{\alpha^2}$.

Proof. From $\tau < \frac{\overline{\gamma}}{\alpha^2}$ we get $\overline{\gamma} - \tau\alpha^2 > 0$ and hence

$$\begin{aligned}\left\langle \tilde{F}(u) - \tilde{F}(v), u - v \right\rangle &= \langle \nabla J(u) - \nabla J(v), u - v \rangle - \tau\alpha^2 |u - v|^2_C \\ &\geq \left(\overline{\gamma} - \tau\alpha^2\right) |u - v|^2_C + \overline{\gamma} |u - v|^2_{C_0} \\ &\geq \underbrace{\left(\overline{\gamma} - \tau\alpha^2\right)}_{>0} |u - v|^2_{C+C_0}.\end{aligned}$$

□

A.5 Properties of Lipschitz Continuous Operators

In order to be able to extract extract properties only from the domain where a Lipschitz continuous operator is smooth, we need the following generalized mean value theorem.

Proposition A.1. *Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be Lipschitz continuous on an open convex set $U \subset \mathbb{R}^n$ and $u, v \in U$. Then one has*

$$T(u) - T(v) \in \langle \operatorname{co} \partial_C T([v, u]), u - v \rangle$$

where $\partial_C T([v, u])$ is the union of all $\partial T(z)$ for all z in the line segment $[v, u]$.

Proof. See [34, Proposition 2.6.5]. □

It is also possible to extract this information from an even smaller subset as the following result shows.

Proposition A.2. *Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be Lipschitz continuous and $\mathcal{S} \subset \mathbb{R}^n$ a set of zero measure. Then*

$$\partial_C T(u) v = \partial_{\mathcal{S}} T(u) v \qquad \forall u \in \mathbb{R}^n,\ v \in \mathbb{R}^m$$

where $\partial_{\mathcal{S}} T$ is defined by

$$\partial_{\mathcal{S}} T(u) := \operatorname{co}\{\lim_{k\to\infty} \nabla T(u_k) : u_k \to u, u_k \in \mathcal{D}_T \setminus \mathcal{S}\}.$$

Proof. See [34, Proposition 2.6.4]. □

Note that it is in general not known if $\partial_C T(x) = \partial_{\mathcal{S}} T(x)$ holds for $m > 1$ (cf. [34, p. 71]). For smooth operators the mean value theorem implies strong monotonicity if the derivatives are uniformly bounded from below. The following lemma generalizes this result to Lipschitz continuous operators requiring boundedness only on an arbitrary dense set where derivatives exists.

Lemma A.3. *Assume that $T : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous on an open convex set U and that*

$$\langle \underline{H}_T v, v\rangle \leq \langle \nabla T(u) v, v\rangle \qquad \forall v \in \mathbb{R}^n, u \in (\mathcal{D}_T \cap U) \setminus \mathcal{S} \tag{A.3}$$

holds with a symmetric positive semidefinite matrix $\underline{H}_T$ and a zero measure exceptional set $\mathcal{S} \subset \mathbb{R}^n$. Then T is strongly monotone on U with respect to the semi-norm introduced by $\underline{H}_T$, i.e.

$$\langle T(u) - T(v), u - v\rangle \geq \langle u - v, u - v\rangle_{\underline{H}_T} \qquad \forall u, v \in U.$$

Proof. First we note that by Proposition A.2 for $x \in U$, $v \in \mathbb{R}^n$ and $Z \in \partial_C T(u)$ there are sequences $u_k^i \in (\mathcal{D}_T \cap U) \setminus \mathcal{S}$ with $u_k^i \to u$ for $k \to \infty$ and a convex combination $\lambda_1, \dots, \lambda_{n^2+1} \geq 0$, $\sum \lambda_i = 1$ sucht that

$$\langle Zv, v\rangle = \left\langle \left(\sum_i \lambda_i \lim_{k\to\infty} \nabla T(u_k^i) \right) v, v \right\rangle = \lim_{k\to\infty} \sum_i \lambda_i \left\langle \nabla T(u_k^i) v, v\right\rangle \geq \langle \underline{H}_T v, v\rangle .$$

By the generalization mean value theorem in Proposition A.1 there are also $\lambda_1, \dots, \lambda_R \geq 0$, $\sum \lambda_i = 1$, $t_1, \dots, t_R \in [0,1]$, and $Z_1, ..., Z_R$ with $Z_r \in \partial_C T(u + t_r(v - u))$ such that

$$T(u) - T(v) = \sum_{i=1}^{R} \lambda_i Z_i (u - v)$$

Together with the above estimate of $\langle Zx, x\rangle$ for $Z \in \partial_C T(y)$ this yields the assertion. □

A.6 Truncated Matrices and the Moore–Penrose Pseudoinverse

In this section we collect properties of so-called truncated matrices obtained by restricting a given matrix to certain subspaces. For convenience we repeat the definition originally given in Chapter 4 here:

Definition A.2. *Let $\mathcal{I}, \mathcal{J} \subset \mathbb{N}$ be index sets, $x \in \mathbb{R}^n$ a vector, and $M \in \mathbb{R}^{m,n}$ a matrix. Then define the truncated matrix $M_{\mathcal{I},\mathcal{J}} \in \mathbb{R}^{m,n}$ and the truncated vector $x_{\mathcal{I}} \in \mathbb{R}^n$ by*

$$(M_{\mathcal{I}\mathcal{J}})_{ij} := \begin{cases} M_{ij} & \text{if } i \in \mathcal{I} \text{ and } j \in \mathcal{J}, \\ 0 & \text{else,} \end{cases} \qquad (x_{\mathcal{I}})_i := \begin{cases} x_i & \text{if } i \in \mathcal{I}, \\ 0 & \text{else.} \end{cases}$$

Furthermore, define the abbreviation $M_{\mathcal{I}} := M_{\mathcal{I},\mathcal{I}}$.

Lemma A.4. *Let $\mathcal{I}, \mathcal{J} \in \mathbb{N}$ be index sets and $M \in \mathbb{R}^{n,m}$ a matrix, and $x \in \mathbb{R}^n$. Then*

$$M_{\mathcal{I},\mathbb{N}} = I_{\mathcal{I}} M, \quad M_{\mathbb{N},\mathcal{I}} = M I_{\mathcal{I}}, \quad M_{\mathcal{I},\mathcal{J}} = I_{\mathcal{I}} M I_{\mathcal{J}}, \quad M(x_{\mathcal{I}}) = M_{\mathbb{N},\mathcal{I}} x, \quad (Mx)_{\mathcal{I}} = M_{\mathcal{I},\mathbb{N}} x.$$

Proof. The identities follow instantly from Definition A.2. □

Although truncated matrices are singular we can still invert the corresponding operators on suitable subspaces and represent the inverse operator as matrix. In order to simplify the notation for these opeartors we recall the definition of the Moore–Penrose pseudoinverse $M^+ \in \mathbb{R}^{n,m}$ of a matrix $M \in \mathbb{R}^{m,n}$ using Tikhonov regularization of the symmetric positive semidefinite matrix MM^T.

Definition A.3. *Let $M \in \mathbb{R}^{m,n}$. Then the Moore–Penrose pseudoinverse $M^+ \in \mathbb{R}^{n,m}$ is given by*

$$M^+ := \lim_{\epsilon \to 0} (M^T M + \epsilon I)^{-1} M^T = \lim_{\epsilon \to 0} M^T (M M^T + \epsilon I)^{-1}$$

of a matrix $M \in \mathbb{R}^{m,n}$.

This limit exists even for singular matrices and M^+ can equivalently be defined as the unique matrix satisfying

$$MM^+M = M, \quad M^+MM^+ = M^+, \quad (MM^+)^T = MM^+, \quad (M^+M)^T = M^+M. \tag{A.4}$$

The pseudoinverse reduced to the inverse if M is invertible. Beside this it has the properties

$$M^{++} = M, \qquad (M^T)^+ = (M^+)^T, \qquad (\lambda M)^+ = \lambda^{-1} M^+ \qquad \forall \lambda > 0.$$

Lemma A.5. *Let $M \in \mathbb{R}^{n,n}$ be symmetric positive definite and $\mathcal{I} \subset \mathbb{N}$ an index set. Then we have*

$$M_{\mathcal{I}}^+ := (M_{\mathcal{I}})^+ = \Big((M_{\mathcal{I}} + I - I_{\mathcal{I}})^{-1} \Big)_{\mathcal{I}} = \Big((M_{\mathcal{I}})^+ \Big)_{\mathcal{I}}.$$

Hence $M_{\mathcal{I}}^+$ is obtained by deleting the i-th rows and columns of M with $i \in \mathcal{I}$, taking the inverse of the reduced matrix and then inserting zero rows and columns where rows and columns where deleted.

Proof. The matrix $(M_{\mathcal{I}} + I - I_{\mathcal{I}})$ is also symmetric positive definite and thus invertible. Its easy to see that for $i, j \in \mathcal{I}$ we have $(M_{\mathcal{I}} + I - I_{\mathcal{I}})^{-1}_{ij} = \delta_{ij}$. Hence we have

$$(M_{\mathcal{I}} + I - I_{\mathcal{I}})^{-1} = \Big((M_{\mathcal{I}} + I - I_{\mathcal{I}})^{-1} \Big)_{\mathcal{I}} + I - I_{\mathcal{I}}.$$

Using this and $(I - I_{\mathcal{I}}) I_{\mathcal{I}} = 0$ we get

$$\Big((M_{\mathcal{I}} + I - I_{\mathcal{I}})^{-1} \Big)_{\mathcal{I}} M_{\mathcal{I}} = M_{\mathcal{I}} \Big((M_{\mathcal{I}} + I - I_{\mathcal{I}})^{-1} \Big)_{\mathcal{I}} = I_{\mathcal{I}}.$$

This implies the four identifies in (A.4). □

The pseudoinverse will also be helpful to handle limits of inverse matrices.

Lemma A.6. *Let (M^k) with $M^k \in \mathbb{R}^{n,n}$ be a sequence of symmetric positive definite matrices such that $M^k \to M$ for some symmetric positive definite matrix $M \in \mathbb{R}^{n,n}$. Furthermore, let $\mathcal{A} \subset \mathbb{N}$ be an index set and $\alpha^k > 0$ with $\alpha^k \to \infty$. Then*

$$\lim_{k\to\infty} (M^k + \alpha^k I_{\mathcal{A}})^{-1} = ((I - I_{\mathcal{A}})M(I - I_{\mathcal{A}}))^+ = (M_{\mathcal{I}})^+$$

with $\mathcal{I} = \mathbb{N} \setminus \mathcal{A}$.

Proof. Without loss of generality we assume that the indices are ordered such that $i < j$ for all $i \in \mathcal{I}$ and $j \in \mathcal{A}$. Furthermore, we identify truncated matrices with the corresponding submatrices where truncated rows and columns are deleted. We will especially identify $I_{\mathcal{A}}$ with I as submatrix. Then we have

$$M^k + \alpha^k I_{\mathcal{A}} = \begin{pmatrix} M^k_{\mathcal{I}} & M^k_{\mathcal{I},\mathcal{A}} \\ M^k_{\mathcal{A},\mathcal{I}} & M^k_{\mathcal{A}} + \alpha^k I \end{pmatrix}.$$

Using the Schur complement

$$S^k := M^k_{\mathcal{A}} + \alpha^k I - M^k_{\mathcal{A},\mathcal{I}} \left(M^k_{\mathcal{I}}\right)^{-1} M^k_{\mathcal{I},\mathcal{A}}$$

the inverse is given by the decomposition

$$\left(M^k + \alpha^k I_{\mathcal{A}}\right)^{-1} =$$
$$\begin{pmatrix} I & -\left(M^k_{\mathcal{I}}\right)^{-1} M^k_{\mathcal{I},\mathcal{A}} \\ 0 & I \end{pmatrix} \begin{pmatrix} \left(M^k_{\mathcal{I}}\right)^{-1} & 0 \\ 0 & (S^k)^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -M^k_{\mathcal{A},\mathcal{I}} \left(M^k_{\mathcal{I}}\right)^{-1} & I \end{pmatrix}. \tag{A.5}$$

From $M^k \to M$ we get that the submatrices converge to the corresponding submatrix of the limit. By continuity of building the inverse of (sub-)matrices we also have

$$\left(M^k_{\mathcal{I}}\right)^{-1} \to (M_{\mathcal{I}})^{-1}.$$

Since all matrices $M^k + \alpha^k I_{\mathcal{A}}$ are symmetric and positive definite for arbitrary α^k it is well known that the same is true for the Schur complements S^k. Thus it is always invertible and we have

$$\left(S^k\right)^{-1} = \frac{1}{\alpha^k} \left[\frac{1}{\alpha^k} \left(M^k_{\mathcal{A}} - M^k_{\mathcal{A},\mathcal{I}} \left(M^k_{\mathcal{I}}\right)^{-1} M^k_{\mathcal{I},\mathcal{A}} \right) + I \right]^{-1}.$$

From the convergence of the submatrices we know that the term containing only submatrices converges to some matrix. Hence we the whole matrix that should be inverted converges to I and from this we get $(S^k)^{-1} \to 0$. Inserting all convergence results for submatrices in the above decomposition we get

$$\left(M^k + \alpha^k I_{\mathcal{A}}\right)^{-1} \to \begin{pmatrix} (M_{\mathcal{I}})^{-1} & 0 \\ 0 & 0 \end{pmatrix} = (M_{\mathcal{I}})^+$$

where we switch back from submatrices to trucated matrices in the last identity. Note that the essential step to take the limit after factoring $1/\alpha^k$ out is not possible for the original full matrix since the remaining sequence $(1/\alpha^k M^k + I_{\mathcal{A}})$ converges to $I_{\mathcal{A}}$ which is not invertible as full matrix. □

B List of Symbols and Notation

Ω	a bounded, open, and nonempty domain in $\mathbb{R}^d$, $d = 1, 2, 3$	5
$\lvert\Omega\rvert$	measure of the set Ω	5
χ_K	indicator functional of convex set K	8
$\nabla^2 f$	Hessian matrix of a functional $f : \mathbb{R}^d \supset M \to \mathbb{R}$	9
$\operatorname{co} M$	convex hull of the set M	11
S^{d-1}	The unit $(d-1)$-sphere in $\mathbb{R}^d$	11
E	Ginzburg–Landau free energy	13
$\langle \cdot, \cdot \rangle$	Euclidean inner product on $\mathbb{R}^d$	14
$\lVert \cdot \rVert$	Euclidean norm on $\mathbb{R}^d$	14
$\langle \cdot, \cdot \rangle_M$	bilinear form $\langle \cdot, \cdot \rangle_M = \langle M \cdot, \cdot \rangle$ on $\mathbb{R}^n$	14
$\lVert \cdot \rVert_M$	(semi-)norm on $\mathbb{R}^n$ induced by $\langle \cdot, \cdot \rangle_M$	14
$L^p(\Omega)$	Lebesque space	14
$\lVert\cdot\rVert_0$	norm on function space $L^2(\Omega)$	14
$(\cdot, \cdot)$	inner product on function space $L^2(\Omega)$	14
$H^k(\Omega)$	Sobolev space with weak derivatives up to order k	14
$\lVert \cdot \rVert_k$	norm of the Sobolev space $H^k(\Omega)$	14
$H^k_0(\Omega)$	Sobolev space with zero boundary values	14
$\lVert\cdot\rVert_V$	norm of a vector space V	14
$(\cdot, \cdot)_H$	inner product of a pre-Hilbert space H	14
V'	dual space of a normed space V	14
$\langle \cdot, \cdot \rangle$	dual pairing of V' and V for a function space V	14
∂	subdifferential of convex function	16
a.e.	almost everywhere	16

$\mathcal{K}$	constrains set for phase field $\mathcal{K} := \{v \in H^1(\Omega) : \lvert v\rvert \leq 1, \text{a.e. in } \Omega\}$	18
$\mathcal{T}$	a (simplicial) triangulation of a domain in $\mathbb{R}^d$	23
$\mathcal{N}(\mathcal{T})$	set of all nodes of $\mathcal{T}$	23
$\mathcal{E}(\mathcal{T})$	set of all edges of $\mathcal{T}$	23
$\mathcal{H}(\mathcal{T})$	set of all hanging nodes of $\mathcal{T}$	25
$(\mathcal{T}_0, \dots, \mathcal{T}_j)$	a grid hierarchy	24
$\mathcal{L}(\mathcal{T}_0, \dots, \mathcal{T}_j)$	leaf grid of a grid hierarchy	24
$\mathcal{S}(\mathcal{T})$	first order conforming finite element space on $\mathcal{T}$	25
$\mathcal{B}(\mathcal{T})$	conforming nodal basis of $\mathcal{S}(\mathcal{T})$	26
$(\cdot, \cdot)^{\mathcal{T}}$	lumped L^2 inner product on $\mathcal{S}(\mathcal{T})$ w.r.t. a triangulation $\mathcal{T}$	29
$\mathcal{D}_T$	set where the Lipschitz continuous operator T is differentiable	48
∂_B	B-subdifferential	48
∂_C	generalized Jacobian in the sense of Clarke	48
∂	generalized linearization of operator	48
∂^2	generalized Hessian of functional	48
$\rho(J, x, d)$	minimizer of J along the line $x + \rho d$ with $\rho \in \mathbb{R}$	52
I	identity matrix on $\mathbb{R}^n$	58
$M_{\mathcal{I},\mathcal{J}}$	matrix with (i, j)-th entry set to zero if $i \in \mathcal{I}$ or $j \in \mathcal{J}$	59
$M_{\mathcal{I}}$	matrix with i-th row and column set to zero if $i \in \mathcal{I}$	59
$x_{\mathcal{I}}$	vector with i-th component set to zero if $i \in \mathcal{I}$	59
M^+	Moore–Penrose pseudoinverse of the matrix M	59
$\mathcal{I}(v)$	set of all indices i where $\partial\varphi_i(v_i)$ is single-valued	60
$\mathcal{I}'(v)$	set of all indices $i \in \mathcal{I}(v)$ where $\varphi_i''(v_i)$ is finite	60
$\mathcal{I}''(v)$	set of all indices $i \in \mathcal{I}'(v)$ where $\varphi_i''(v_i)$ is bounded by C_φ	63
PDAS	primal–dual active set method	66
TNNMG	truncated nonsmooth Newton multigrid method	72
$[c]$	set of all v with the same active configuration as c	92

$\mathcal{C}$	set of all active configurations of φ	92
$\mathcal{E}$	set of all multi-indices of intervals where φ is smooth	93
$[c]_\eta$	set of all $v \in [c]$ with v_i in interior of η_i-th smooth interval	93
$[[c]]_\eta$	set of all $v \in [c]$ with v_i in closure of η_i-th smooth interval	93
$F([c]_\eta)^\circ$	dense open subset of $F([c]_\eta)$	93

C List of Assumptions

Bibliography

[1] M. Ainsworth and J. T. Oden. *A posteriori error estimation in FE analysis.* Wiley, 2000.

[2] S. Allen and J. Cahn. A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. *Acta Metall.*, 27:1084–1095, 1979.

[3] J. Appell and P. P. Zabrejko. *Nonlinear Superposition Operators.* Number 95 in Cambridge Tracts in Mathematics. Cambridge Univiversity Press, Cambridge, 1990.

[4] L. Armijo. Minimization of functions having Lipschitz–continuous first partial derivatives. *Pazific J. Math.*, 204:126–136, 1966.

[5] L. Badea. Convergence rate of a Schwarz multilevel method for the constrained minimization of nonquadratic functionals. *SIAM J. Numer. Anal.*, 44(2):449–477, 2006.

[6] L. Badea, X.-Ch. Tai, and J. Wang. Convergence rate analysis of a multiplicative Schwarz method for variational inequalities. *SIAM J. Numer. Anal.*, 41(3):1052–1073, 2003.

[7] J. W. Barrett and J. F. Blowey. Finite element approximation of an Allen–Cahn/Cahn–Hilliard system. *IMA Journal of Numerical Analysis*, 22(1):11–71, 2002.

[8] J. W. Barrett, J. F. Blowey, and H. Garcke. Finite element approximation of the Cahn-Hilliard equation with degenerate mobility. *SIAM J. Numer. Anal.*, 37(1): 286–318, 1999.

[9] J. W. Barrett, R. Nürnberg, and V. Styles. Finite element approximation of a phase field model for void electromigration. *SIAM J. Numer. Anal.*, 42(2): 738–772, 2004.

[10] P. Bastian, K. Birken, S. Lang, K. Johannsen, N. Neuß, H. Rentz-Reichert, and C. Wieners. UG: A flexible software toolbox for solving partial differential equations. *Computing and Visualization in Science*, 1:27–40, 1997.

[11] P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, R. Kornhuber, M. Ohlberger, and O. Sander. A generic interface for adaptive and parallel scientific computing. Part II: Implementation and tests in DUNE. *Computing*, 82(2–3):121–138, 2008.

[12] P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, M. Ohlberger, and O. Sander. A generic interface for adaptive and parallel scientific computing. Part I: Abstract framework. *Computing*, 82(2–3):103–119, 2008.

[13] G. Bellettini and M. Paolini. Anisotropic motion by mean curvature in the context of finsler geometry. *Hokkaido Math. J.*, 25:537–566, 1996.

[14] M. Bergounioux, K. Ito, and K. Kunisch. Primal–dual strategy for constrained optimal control problems. *SIAM J. Control Optim.*, 37:1176–1194, 1999.

[15] J. F. Blowey and C. M. Elliot. The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy part I: Mathematical analysis. *European J. Appl. Math.*, 2:233–280, 1991.

[16] J. F. Blowey and C. M. Elliot. The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy part II: Numerical analysis. *European J. Appl. Math.*, 3:147–179, 1992.

[17] T. Böhme, W. Dreyer, F. Duderstadt, and W. H. Müller. A higher gradient theory of mixtures for multi-component materials with numerical examples for binary alloys. Preprint 1286, WIAS, 2007.

[18] F. A. Bornemann. *An Adaptive Multilevel Approach for Parabolic Equations in Two Space Dimensions.* PhD thesis, Freie Universität Berlin, 1991.

[19] F. A. Bornemann, B. Erdmann, and R. Kornhuber. A posteriori error estimates for elliptic problems in two and three space dimensions. *SIAM J. Numer. Anal.*, 33:1188–1204, 1996.

[20] D. Braess and W. Hackbusch. A new convergence proof for the multigrid method including the V–cycle. *SIAM J. Numer. Anal.*, 20:967–975, 1983.

[21] D. Braess and R. Sarazin. An efficient smoother for the Stokes problem. *Appl. Numer. Math.*, 23(1):3–19, 1997.

[22] J. H. Bramble, J. E. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Math. Comp.*, 55(191):1–22, 1990.

[23] J. H. Bramble, J. E. Pasciak, Wang J., and J. Jinchao Xu. Convergence estimates for multigrid algorithms without regularity assumptions. *Math. Comp.*, 57(195): 23–45, 1991.

[24] J. H. Bramble, J. E. Pasciak, Wang J., and J. Jinchao Xu. Convergence estimates for product iterative methods with applications to domain decomposition. *Math. Comp.*, 57(195):1–21, 1991.

[25] M. Brokate and J. Sprekels. *Hysteresis and Phase Transition.* Number 121 in Applied mathematical sciences. Springer, Berlin Heidelberg New York, 1996.

[26] E. Burman and J. Rappaz. Existence of solutions to an anisotropic phase-field model. *Math. Meth. App. Sci.*, 26:1137–1160, 2003.

[27] A. Burri, A. Dedner, R. Klöfkorn, and M. Ohlberger. An efficient implementation of an adaptive and parallel grid in dune. In *Proc. of the 2nd Russian-German Advanced Research Workshop on Computational Science and High Performance Computing*, 2005.

[28] G. Caginalp. An analysis of a phase field model of a free boundary. *Arch. Rat. Mech. Anal.*, 92:205–245, 1986.

[29] J. W. Cahn and J. E. Hilliard. Free energy of a non-uniform system i. interfacial free energy. *Jnl. of Chemical Physic*, 28:258–267, 1958.

[30] J. W. Cahn and J. E. Taylor. Linking anisotropic sharp and diffuse surface motion laws via gradient flows. *Journal of Statistical Physics*, 77(1–2):187–197, 1994.

[31] J. W. Cahn and J. E. Taylor. Overview no. 113 surface motion by surface diffusion. *Acta Metallurgica et Materialia*, 42(4):1045–1063, 1994.

[32] J. W. Cahn, C. M. Elliott, and A. Novick-Cohen. The Cahn–Hilliard equation with concentration dependent mobility: motion by minus the Laplace of the mean curvature. *European J. Appl. Math.*, 7:287–301, 1996.

[33] X. Chen and C. M. Elliott. Asymptotics for a parabolic double obstacle problem. *Proc. R. Soc. Lond., Ser. A*, 444:429–445, 1994.

[34] F. H. Clarke. *Optimization and Nonsmooth Analysis.* Wiley, New York, 1983.

[35] M. I. M Copetti and C. M. Elliott. Numerical analysis of the Cahn-Hilliard equation with a logarithmic free energy. *Numerische Mathematik*, 63:39–65, 1992.

[36] R. W. Cottle, J. S. Pang, and R. E. Stone. *The Linear Complentary Problem.* Academic Press, Boston, 1992.

[37] M. G. Crandall and P.-L. Lions. Viscosity solutions of Hamilton–Jacobi equations. *Trans. Amer. Math. Soc.*, 277(1):1–42, 1983.

[38] R. Dautray and J. L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology. 5: Evolution Problems.* Springer, Berlin Heidelberg New York, 1992.

[39] T. A. Davis. Algorithm 832: Umfpack v4.3 – an unsymmetric-pattern multifrontal method. *ACM Trans. Math. Softw.*, 30(2):196–199, 2004.

[40] P. de Mottoni and M. Schatzman. Geometrical evolution of developed interfaces. *Trans. Am. Math. Soc.*, 347(5):1533–1589, 1995.

[41] K. Deckelnick, G. Dziuk, and C. M. Elliott. Computation of geometric partial differential equations and mean curvature flow. *Acta Numer.*, 14:139–232, 2005.

[42] J. W. Demmel, S. C. Eisenstat, J. R. Gilbert, X. S. Li, and J. W. H. Liu. A supernodal approach to sparse partial pivoting. *SIAM J. Matrix Analysis and Applications*, 20(3):720–755, 1999.

[43] P. Deuflhard. *Newton Methods for Nonlinear Problems.* Number 35 in Springer Series in Computational Mathematics. Springer, Berlin Heidelberg New York, 1. edition, 2004.

[44] P. Deuflhard and M. Weiser. *Numerische Mathematik 3.* Walter de Gruyter, Berlin, 2011.

[45] P. Deuflhard, P. Leinen, and H. Yserentant. Concepts of an adaptive hierarchical finite element code. *IMPACT Comput. Sci. Engrg.*, 1:3–35, 1989.

[46] W. Dörfler. A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.*, 33:1106–1124, 1996.

[47] W. Dörfler and R. H. Nochetto. Small data oscillation implies the saturation assumption. *Numer. Math.*, 91:1–12, 2002.

[48] C. Eck, H. Garcke, and P. Knabner. *Mathematische Modellierung.* Springer, Berlin Heidelberg New York, 2008.

[49] I. Ekeland and R. Temam. *Convex Analysis.* North-Holland, 1976.

[50] C. M. Elliott. The Cahn-Hilliard model for the kinetics of phase separation. In J. F. Rodrigues, editor, *Mathematical Models for Phase Change Problems*, volume 88 of *International Series of Numerical Mathematics.* Birkhäuser, Basel, 1989.

[51] C. M. Elliott and S. Luckhaus. A generalised diffusion equation for phase separation of a multicomponent mixture with interfacial free energy. Preprint 195, University of Bonn, 1991.

[52] C. M. Elliott and R. Schätzle. The limit of the fully anisotropic double-obstacle Allen–Cahn equation in the nonsmooth case. *SIAM J. Math. Anal.*, 28(2):274–303, 1997.

[53] C. Geiger and C. Kanzow. *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben.* Springer, Berlin Heidelberg New York, 1999.

[54] R. Glowinski. *Numerical Methods for Nonlinear Variational Problems.* Springer Series in Computational Physics. Springer, Berlin Heidelberg New York, 3. edition, 1984.

[55] R. Glowinski, J. L. Lions, and R. Trémolières. *Numerical Analysis of Variational Inequalities.* Number 8 in Studies in Mathematics and its Applications. North-Holland Publishing Company, Amsterdam New York Oxford, 1981.

[56] C. Gräser. Analysis und Approximation der Cahn-Hilliard Gleichung mit Hindernispotential. Diplomarbeit, Freie Universität Berlin, 2004.

[57] C. Gräser. Globalization of nonsmooth Newton methods for optimal control problems. In K. Kunisch, G. Of, and O. Steinbach, editors, *Numerical Mathematics and Advanced Applications*, number 60 in LNCSE, pages 605–612, Berlin, 2007. Springer.

[58] C. Gräser and R. Kornhuber. Multigrid methods for obstacle problems. *J. Comp. Math.*, 27(1):1–44, 2009.

[59] C. Gräser and R. Kornhuber. Nonsmooth Newton methods for set-valued saddle point problems. *SIAM J. Numer. Anal.*, 47(2):1251–1273, 2009.

[60] C. Gräser and O. Sander. The dune-subgrid module and some applications. *Computing*, 8(4):269–290, 2009.

[61] C. Gräser, R. Kornhuber, and U. Sack. On hierarchical error estimators for time-discretized phase field models. In *Proceedings of ENUMATH 2009*, 2009. accepted.

[62] C. Gräser, U. Sack, and O. Sander. Truncated nonsmooth Newton multigrid methods for convex minimization problems. In M. Bercovier, M. Gander, R. Kornhuber, and O. Widlund, editors, *Domain Decomposition Methods in Science and Engineering XVIII*, LNCSE, pages 129–136. Springer, 2009.

[63] W. Hackbusch. *Multi-Grid Methods and Applications.* Springer, Berlin, 1985.

[64] 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.

[65] R. H. W. Hoppe. Multigrid algorithms for variational inequalities. *SIAM J. Numer. Anal.*, 24:1046–1065, 1987.

[66] R. H. W. Hoppe and R. Kornhuber. Adaptive multilevel–methods for obstacle problems. *SIAM J. Numer. Anal.*, 31(2):301–323, 1994.

[67] K. Ito and K. Kunisch. Convergence of the primal–dual active set strategy for diagonally dominant systems. *SIAM J. Control Optim.*, 13(1):14–34, 2007.

[68] R. Kobayashi. Modeling and numerical simulations of dendritic crystal growth. *Physica D*, 63:410 – 423, 1993.

[69] R. Kornhuber. Monotone multigrid methods for elliptic variational inequalities I. *Numer. Math.*, 69:167 – 184, 1994.

[70] R. Kornhuber. A posteriori error estimates for elliptic variational inequalities. *Comput. Math. Appl.*, 31:49–60, 1996.

[71] R. Kornhuber. Monotone multigrid methods for elliptic variational inequalities II. *Numer. Math.*, 72:481 – 499, 1996.

[72] R. Kornhuber. *Adaptive Monoton Multigrid Methods for Nonlinear Variational Problems.* Teubner, Stuttgart, 1. edition, 1997.

[73] R. Kornhuber. On constrained Newton linearization and multigrid for variational inequalities. Preprint A/02/2001, FU Berlin, 2001.

[74] R. Kornhuber. Nonlinear multigrid techniques. In J. F. Blowey, J. P. Coleman, and A. W. Craig, editors, *Theory and Numerics of Differential Equations*, pages 179–229, Heidelberg, 2001. Springer.

[75] R. Kornhuber. On constrained Newton linearization and multigrid for variational inequalities. *Numer. Math.*, 91:699–721, 2002.

[76] R. Kornhuber and R. Krause. Robust multigrid methods for vector-valued Allen–Cahn equations with logarithmic free energy. *Comp. Visual. Sci.*, 9:103–116, 2006.

[77] R. Kornhuber and H. Yserentant. Multilevel methods for elliptic problems on domains not resolved by the coarse grid. *Contemp. Math.*, 180:49–60, 1994.

[78] R. Kornhuber and Q. Zou. Efficient and reliable hierarchical error estimates for the discretization error of elliptic obstacle problems. Preprint 519, Matheon Berlin, 2008.

[79] B. Kummer. Newton's method based on generalized derivatives for nonsmooth functions: Convergence analysis. In W. Oettli and D. Pallaschke, editors, *Advances in optimization (Lambrecht, 1991)*, pages 171–194, Berlin, 1992. Springer.

[80] J. Mandel. A multilevel iterative method for symmetric, positive definite linear complementarity problems. *Appl. Math. Optimization*, 11:77–95, 1984.

[81] A. Nekvinda and L. Zajíček. A simple proof of the Rademacher theorem. *Časopis Pěst. Mat*, 113(4):337–341, 1988.

[82] J. Nocedal. Theory of algorithms for unconstrained optimization. *Acta Numerica*, 1:199–242, 1992.

[83] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, New York, 1970.

[84] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM J. Numerical Analysis*, 12:617–629, 1975.

[85] J. S. Pang. Newton's method for b-differentiable equations. *Mathematics of Operations Research*, 15(2):311–341, 1990.

[86] O. Penrose and P. C. Fife. Thermodynamically consistent models of phase field type for the kinetics of phase transitions. *Physica D*, 43:44–62, 1990.

[87] M. J. D. Powell. Direct search algorithms for optimization calculations. *Acta Numerica*, 7:287–336, 1998.

[88] L. Qi. Convergence analysis of some algorithms for solving nonsmooth equations. *Mathematics of Operations Research*, 18(1):227–244, 1993.

[89] L. Qi and J. Sun. A nonsmooth version of Newtons's method. *Mathematical Programming*, 58:353–367, 1993.

[90] A. Rätz, A. Ribalta, and A. Voigt. Surface evolution of elastically stressed films under deposition by a diffuse interface model. *J. Comp. Phys.*, 214(1):187–208, 2006.

[91] J. Rubinstein, P. Sternberg, and J. B. Keller. Fast reaction, slow diffusion and curve shorteneing. *SIAM J. Appl. Math.*, 49:116–133, 1989.

[92] J. W. Ruge and K. Stüben. Algebraic multigrid. In S. F. McCormick, editor, *Multigrid Methods*, pages 73–130. SIAM, Philadalphia, 1987.

[93] Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comp.*, 7:856–869, 1986.

[94] O. Sander. *Multidimensional Coupling in a Human Knee Model.* PhD thesis, Freie Universität Berlin, 2008.

[95] A. Schiela. *The Control Reduced Interior Point Method. A Function Space Oriented Algorithmic Approach.* PhD thesis, Freie Universität Berlin, 2006.

[96] A. Schiela. A simplified approach to semismooth Newton methods in function space. *SIAM J. Control Optim.*, 19(3):1417–1432, 2008.

[97] A. Schmidt and K. G. Siebert. *Design of Adaptive Finite Element Software. The Finite Element Toolbox ALBERTA*, volume 42 of *LNCSE.* Springer, Berlin Heidelberg New York, 2005.

[98] J. Schöberl and W. Zulehner. On Schwarz-type smoothers for saddle point problems. *Numer. Math.*, 95(2):377–399, 2003.

[99] B. Schupp. *Entwicklung eines effizienten Verfahrens zur Simulation kompressibler Strömungen in 3D auf Parallelrechnern.* PhD thesis, Albert-Ludwigs-Universität Freiburg, 1999.

[100] K. G. Siebert and A. Veeser. A unilaterally constrained qadratic minimization with adaptive finite elements. *SIAM J. Optim.*, 18:260–289, 2007.

[101] R. Simon and W. Zulehner. On Schwarz-type smoothers for saddle point problems with applications to PDE-constrained optimization problems. *Numer. Math.*, 111:445–468, 2009.

[102] P. Spellucci. *Numerische Verfahren der nichtlinearen Optimierung.* Birkhäuser, Basel Berlin, 1993.

[103] B. Stinner. *Derivation and Analysis of a Phase Field Model for Alloy Solidification.* PhD thesis, Universität Regensburg, 2005.

[104] X.-C. Tai. Rate of convergence for some constraint decomposition methods for nonlinear variational inequalities. *Numer. Math.*, 93(4):755–786, 2003.

[105] S. Torabi, S. Wise, J. Lowengrub, A. Rätz, and A. Voigt. A new method for simulating strongly anisotropic Cahn-Hilliard equations. In *Materials Science and Technology-Association for Iron and Steel Technology*, volume 3, pages 1432–1444. Curran Associates, Inc., 2007.

[106] M. Ulbrich. *Nonsmooth Newon-like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces.* Habilitationsschrift, Technische Universität München, 2002.

[107] S. P. Vanka. Block-implicit multigrid solution of Navier-Stokes equations in primitive variables. *J. Comput. Phys.*, 65:138–158, 1986.

[108] D. Werner. *Funktionalanalysis.* Springer, Berlin Heidelberg New York, 3. edition, 2000.

[109] S. Wise, J. Kim, and J. Lowengrub. Solving the regularized, strongly anisotropic Cahn-Hilliard equation by an adaptive nonlinear multigrid method. *J. Comp. Phys.*, 226(1):414–446, 2007.

[110] J. Wloka. *Partielle Differentialgleichungen.* B. G. Teubner, Stuttgart, 1982.

[111] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Review*, 34:581–613, 1992.

[112] H. Yserentant. Two preconditioners based on the mutilevel splitting of finite element spaces. *Numer. Math.*, 58:163–184, 1990.

[113] H. Yserentant. Old and new convergence proofs for multigrid methods. *Acta Numerica*, pages 285–326, 1993.

[114] O. C. Zienkiewicz, J. P. De S. R. Gago, and D. W. Kelly. The hierarchical concept in finite element analysis. *Computers & Structures*, 16:53–65, 1983.

[115] Q. Zou, A. Veeser, R. Kornhuber, and C. Gräser. Hierarchical error estimates for the energy functional in obstacle problems. *Numer. Math.*, 2009. submitted.

[116] W. Zulehner. A class of smoothers for saddle point problems. *Computing*, 65(3): 227–246, 2000.

[117] W. Zulehner. Analysis of iterative methods for saddle point problems: A unified approach. *Math. Comput.*, 71(238):479–505, 2002.

Zusammenfassung

Phasenfeldmodelle sind ein weit verbreiteter Ansatz zur Beschreibung von Prozessen, die sich wesentlich durch dünne Interfaceregionen zwischen weitestgehend homogenen Bereichen auszeichnen. Ein wichtiges Anwendungsfeld von Phasenfeldmodellen ist die Modellierung von physikalischen Phasenübergangs- und Phasenseparationvorgängen. Eine wesentliche Eigenschaften ist dabei, dass die Trennung der Phasen durch ein Doppelmuldenpotential getrieben ist, welches voneinander getrennte Minima für jede Phase besitzt. Bereits 1958 haben Cahn und Hilliard ein logarithmisches Potential vorgeschlagen, das zwar differenzierbar ist, aber singuläre Ableitungen besitzt. Geht die Temperatur gegen 0, so degeneriert das temperaturabhängige logarithmische Potential gegen das nichtdifferenzierbare Hindernispotential.

Ziel der vorliegenden Arbeit ist die Entwicklung von Methoden zur effizienten numerischen Lösung solcher Gleichungen, die auch im Fall nichtglatter Potentiale und anisotroper Oberflächenenergien robust sind. Diese Methoden werden für die Cahn-Hilliard-Gleichung entwickelt, die prototypisch für eine Vielzahl solcher Modelle ist.

Das Hauptresultat der Arbeit ist die Entwicklung eines schnellen iterativen Verfahrens zur Lösung nichtlinearer Sattelpunktprobleme, wie sie bei der Diskretisierung anisotroper Cahn-Hilliard-Gleichungen mit Finite-Elemente-Methoden entstehen. Die Grundlage dieses Verfahrens ist eine Umformulierung des Sattelpunktproblems als äquivalentes duales Minimierungsproblem. Das Energiefunktional dieses Minimierungsproblems ist differenzierbar und seine Ableitung ist das nichtlineare Schur-Komplement des Sattelpunktproblems.

Für dieses Schur-Komplement wird eine verallgemeinerte Linearisierung hergeleitet, die im Rahmen eines nichtglatten Newton-Verfahrens Verwendung findet. Für dieses so genannte „Schur Nonsmooth Newton“ Verfahren wird die globale Konvergenz mittels der Äquivalenz zu Abstiegsverfahren für das duale Minimierungsproblem gezeigt. Ferner wird bewiesen, dass die globale Konvergenz auch bei inexaktem Lösen der linearen Newton Probleme erhalten bleibt.

In jedem Schritt dieses Verfahrens ist ein nichtlineares konvexes Minimierungsproblem zu lösen. Für die effiziente Behandlung dieser Teilprobleme wird das so genannte „Truncated Nonsmooth Newton Multigrid“ (TNNMG) Verfahren entwickelt. Dieses nichtlineare Mehrgitterverfahren zeichnet sich im Gegensatz zu verwandten Verfahren durch seine Einfachheit und die Anwendbarkeit auf anisotrope Probleme aus. Hinsichtlich der Konvergenzgeschwindigkeit ist es bereits bekannten Verfahren ebenbürtig.

Numerische Beispiele zeigen, dass die entwickelten Verfahren gitterunabhängig konvergieren. Auch erweisen sich die Verfahren als robust für verschiedene Temperaturen, einschließlich dem Grenzfall 0. Der Grund für diese Robustheit ist, dass die Verfahren nicht auf Differenzierbarkeit sondern der konvexen Struktur der Probleme basieren.